VOLUME 61
NUMBER 2

WHOLE NO. 283
1947

Psychological Monographs

JOHN F. DASHIELL, *Editor*

The Use of Aptitude Tests in the Selection of Radio Tube Mounters

By
LOUIS VINCENT SURGENT

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Faculty of Philosophy, Columbia University.

Published by
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Publications Office
1515 MASSACHUSETTS AVE., N.W., WASHINGTON 5, D.C.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
I. INTRODUCTION AND STATEMENT OF THE PROBLEM	1
II. JOB ANALYSIS	3
A. The Operations Performed	3
B. The Aptitudes Required	9
C. Production Data Examined as Possible Criteria	10
III. THE CORRELATED VARIABLES	13
A. The Criterion	13
B. The Aptitude Tests	15
IV. EXPERIMENTAL CONDITIONS AND SUBJECTS	18
V. RESULTS AND DISCUSSION	20
A. The Criterion Scores	20
B. The Test Scores	23
C. The Correlations between Tests and Criterion	24
D. The Composition and Yield of Selected Test Batteries	25
E. Converting Predictions on the Assumed Scale to Grades on the Rating Scale	27
F. Chart for Reporting Test Performance	28
G. Predicting Factory Performance	29
H. Graphical Presentation of the Results	32
I. The Experience Hypothesis	35
VI. SUMMARY AND CONCLUSIONS	38
BIBLIOGRAPHY	40

ACKNOWLEDGMENTS

INDUSTRIAL personnel research is, by its very nature, a cooperative venture. Although, for reasons of efficiency, a specialist may be charged with responsibility for the planning and execution of a program, the entire process from the shaping of the initial plans to the application of the final results, reflects the understanding, the cooperation and the willing efforts of many, at all levels in the organization. Even the skeptic, with his many and persistent questions, can claim a share in the final product. He is the stone upon which the experimenter sharpens his tools.

In these circumstances, it is difficult to single out a few for special mention. Yet there are those whose contributions are too substantial to acknowledge anonymously.

The author wishes to express appreciation to the management of the Harrison, New Jersey, Plant of the Radio Corporation of America, for its understanding of the importance of proper placement methods and their bearing on manpower utilization during the war, and for its willingness to collect data in scientific fashion and to seek guidance in the results.

To Miss Helen Bircher, who gave so freely of her experience with an understanding of employment problems, and to interviewers, Miss Dorothy Cornthwiate and Mr. James J. Bryant (now Supervisor of Wage and Salary Administration at another branch of the company), the author's debt is manifold. Despite many day to day inconveniences, they continued to supply the subjects for this and other studies. The spirit in which this was done and the encouragement derived from the knowledge that the results of our efforts would find intelligent and understanding application in their hands, were equally important to this experimenter.

It is no exaggeration to say that the present study would not have been possible without the conscientious efforts and continued cooperation of Miss Florence Holton, Supervisor of the Vestibule Training School during the period of these experiments, and her staff which included instructors, Miss Mary Dudak, Miss Helen Stocki and Miss Eleanor Jankowski, and inspector, Miss Helen Chambers. It was they, together with many supervisors in the factory, who provided the data which enabled the validation of the tests.

For the administration of the tests to the majority of applicants and for assistance in the compilation of data and a portion of its statistical treatment, the author wishes to thank his present and former assistants, Miss Ann Shults and Miss Dolores Wargo.

Apologies as well as thanks are due the many applicants who took the tests, unaware of the fact that the results would have no effect upon their final placement.

The author was fortunate in having the advice and guidance of the members of the graduate seminar group at Columbia University, and is especially indebted to Professors H. E. Garrett and A. T. Poffenberger, and to Dr. W. N. Schoenfeld.

The many others who have cooperated in this project will recognize their contributions in the following pages and will know that they have been appreciated.

THE USE OF APTITUDE TESTS IN THE SELECTION OF RADIO TUBE MOUNTERS

I. Introduction and Statement of the Problem

In April, 1942, a Vestibule Training School was organized at the Harrison, New Jersey, Plant of the Radio Corporation of America, in order to train selected new employees quickly in the basic skills involved in the assembly of radio tube mounts and to place graduates in the factory in accordance with the level of ability demonstrated during the training period.

The instructors soon discovered that individuals differed markedly in ability to perform the operations used for training purposes in the school. Specifically, it appeared that the existing employment procedures did not provide for a sufficiently precise evaluation of the manipulative skills required for successful performance.[1]

These and other facts were taken to indicate the desirability of determining experimentally the value of certain manipulative aptitude tests in the selection of trainees for the Vestibule Training School.

After two preliminary experiments in the school and one with experienced mounters in the factory, the follow-up method was adopted and used almost exclusively. The program called for testing prospective mounters at the time of hiring, and for later correlation of the scores with criteria of performance in the school and in the factory.

It was the original intention of this experimenter to present all the data procured in the course of the research with mounters. The approach to the problem, however, was one of continuous and, somewhat later, intermittent research. A number of experimental tests were administered even after the initial battery found application in the employment office. Furthermore, many of the decisions affecting the research with mounters were determined by the possibilities and

[1] The reader who is familiar with even a small part of the literature and research concerning the nature and magnitude of individual differences and the difficulties frequently involved in their detection and evaluation, will not interpret these observations as a reflection upon the competence of those charged with responsibility for selection and placement.

Griffitts (7), for example, studied the relationship between twelve anatomical measurements, including height, weight, hand measurements and various ratios thereof, and performance on a total of twelve manipulative tasks. Despite the fact that the measures of manipulative ability had an average reliability of .85, all correlations between performance and the anthropometric measurements were low. Only two of the 144 coefficients "may . . . have statistical significance." Tiffin (18), commenting upon this and other studies, says in part:

> Some employment managers judge the dexterity of an applicant by examining his hands and fingers. . . . Perhaps in extreme cases, where an applicant has fingers that are stiff or very stubby, one could predict from an examination of his hands that he would probably be low in finger dexterity; but in the great majority of cases such a judgment would be no more than a guess. What an applicant can do with his hands, not the appearance of the hands, determines his qualifications for a manual dexterity job.

Moreover, there is ample evidence to show not only that the correlations between various measures of dexterity and intelligence are generally low (3), (12), (13), (20), but also that dexterity is more or less unrelated to other abilities. Thus Harrell (9) reports that "manual agility" is a factor separate from "perception of detail," "verbal relations," "visualizing spatial relations" and "youth." This conclusion is based upon an analysis of the scores on thirty-four variables including various manual, spatial and verbal tests and such personal data as age, school grade completed, experience on mechanical jobs and supervisory ratings. Ninety-one cotton mill machine-fixers served as subjects.

It appears, therefore, that interviewers should not be expected to evaluate the manipulative abilities of applicants except, perhaps, when a reliable report on past performance in similar or identical work is available to demonstrate possession of an adequate amount of the required skills.

practical demands of the situation in which the work was done, as well as by the broader testing program which was being developed simultaneously. It is necessary, therefore, to select for presentation only a portion of the data pertinent to the use of aptitude tests in the selection of mounters.

We will concern ourselves with the study of 233 female trainees who were tested and hired between May 17th and September 15th, 1943, and who satisfy certain other criteria itemized in Chapter IV of this report.

Utilizing the Wherry-Doolittle method, two regression equations will be developed to predict performance in the Vestibule Training School. The predictions made by these equations will be further validated against subsequent performance on mounting jobs in the factory.

While these data do not constitute the total amount of validation on these tests, the data presented and the conclusions reached represent the more important practical results which emerged from the research up to that time.

A review of the literature indicates that to date only one study dealing with the selection of mounters has been published. Forlano and Kirkpatrick (2), working with twenty mounters from another plant of the Radio Corporation of America, found "A composite of intelligence and personality scores . . . to be effective in predicting the subsequent success of new tube mounters."

The present report will be confined to an evaluation of five standard manipulation tests for this situation.

II. Job Analysis

The assembly of mounts for the receiving and allied type tubes which are manufactured at the Harrison Plant, involves the careful manipulation and positioning of small and often delicate parts with the fingers and tweezers. A very high percentage of these jobs require the use of a bench type, resistance welder. When the Vestibule Training School was organized, therefore, three relatively simple assemblies were selected for fabrication by the process of resistance welding, on the grounds that they embodied skills which are basic to most mounting operations. In order to provide a situation favorable to the rapid acquisition of these skills and, at the same time, to measure the relative ability of each trainee to acquire them, the procedures followed resembled those employed in the administration of aptitude tests of the manipulative type. After a period of instruction and supervised practice with a particular assembly, the trainees were given a series of *production tests* of one hour duration. At the conclusion of each test, the number of units produced was recorded and the work inspected. While there were a number of factors (pp. 11f) in the situation which precluded the use of these data as criteria, it will be shown that they provided the instructors with a reasonably sound basis for evaluating trainee performance.

In order that the operational significance of the experimental results may be understood by the reader who is not familiar with the process of assembling small metal parts by resistance welding, the process, the equipment and the operations performed will be described in some detail. Results of the training procedure will then be examined for their use in selection.

A. The Operations Performed[2]

1. *Resistance Welding: Equipment and Process.*

The joining of two metal parts by resistance welding is accomplished by the application of a potential difference in such a way that an electric current passes from one part to the other at the point (small area) where the weld is desired. The heat developed is proportional to the resistance of the path, the square of the current and its action time, which depend in turn upon other factors, such as the mechanical pressure exerted on the parts and the properties of the materials themselves. Under proper conditions, the current encounters maximum resistance in passing from one part to the other. If sufficient, but not excessive, heat is developed at this point, the parts will be welded. Resistance welding is complicated by the difficulties involved in localizing and controlling the fusion of the two metals in the desired area and in avoiding burning (excessive fusion and oxidation) or welding of the parts to the electrodes. Most of these problems are the concern of engineers, supervisors and setup men and will not be discussed. Of immediate interest are the operations performed and the process factors controlled by the trainee.

Figure 1 shows a bench type resistance welder, mounted on table (T). Two crossed lengths of wire (X) and (Y), are positioned for welding between the stationary electrode (A) and the movable

[2] The author wishes to thank Mr. J. R. Gates, welding engineer, who reviewed this section of the manuscript for technical accuracy. Over a period of several years, numerous supervisors and machine attendants have given freely of their experience with welding problems. In this connection, thanks are especially due to Mr. F. J. Pilas.

electrode (B). If the drawing were complete, it would show an operator seated in an adjustable posture chair, holding wire (X) with a pair of tweezers grasped in the right hand, and wire (Y) in the fingers of the left hand. The operator's feet would be located on the left (L) and right (R) pedals, respectively. The pedal (R) controls the movement of the electrode (B) by means of a series of mechanical linkages which are represented in highly conventionalized form in the sketch. The chain (F), which is attached to the pedal (R), passes over pulley (H) to the lever arm (K). Although (K) is mounted independently of the swinging arm (D) on the shaft (M), the two parts are connected by the spring (E). The drawing shows the right pedal depressed just enough to bring the movable electrode into contact with the work. Without pausing, the operator continues the downward stroke of the right pedal beyond this point until the movement of (K) is arrested by the stop (G). Since the two lengths of metal wire are relatively incompressible when cold, depression of the pedal results in the extension of spring (E). In terms of the conventionalized drawing, the exact pressure exerted on the parts at the moment of welding can be controlled by selecting and ad-

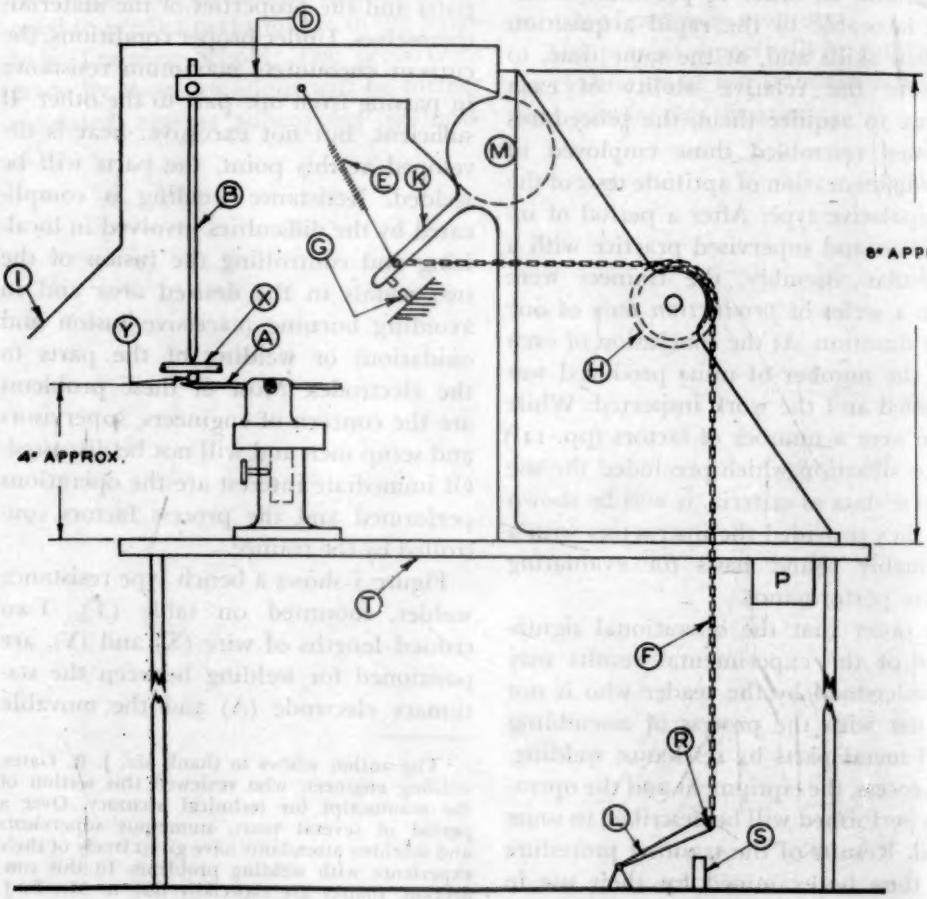


FIG. 1. Conventional Drawing of Bench Type Resistance Welder.

justing the spring (E) and by setting the switch (not shown), which initiates the current flow between the electrodes, so that it closes when the proper spring tension is reached. The action time of the current is determined with a high degree of precision by the electronic control panel (P) and does not depend upon the speed with which the pedal (R) is depressed or upon the quick withdrawal of the foot.

Thus depression of pedal (R) serves to bring the movable electrode (B) into contact with the parts to be welded and, in addition, closes the circuit which provides the current needed to complete the weld.

The left pedal has only one function. It is used for welds requiring current of different (usually greater) intensity. Some welds can be made with the right pedal alone, while others require that the left pedal be depressed before the right pedal is used to bring down the movable electrode and complete the weld. Incorrect use of the pedals may result in a cold or burned weld, which can be detected usually by visual inspection. Since mounting jobs generally require a combination of welds at the two intensity levels, operators must follow rigorously the prescribed patterns.

2. *Process Factors Controlled by the Operator.*

The operator is responsible for producing good welds at a rate determined by time study techniques for each specific job, and by strict adherence to instructions pertaining to the use of materials and equipment. In part, this responsibility involves the following:

a. Summoning the machine attendant or supervisor when welds are not coming through properly. An experienced operator must be able to distinguish a good or satisfactory weld from a poor one.
b. Following directions pertaining to the correct use of foot-pedals, as explained above.
c. Avoiding downward pressure on the stationary electrode (A). Since the operator holds the parts with the fingers or tweezers at a point somewhat removed from the area of welding, any downward pressure on either of the parts to be welded may result in "bowing" the parts, especially if the pressure is continued while the parts are being fused and joined.
d. The careful handling of parts. The majority of radio tube parts must be handled lightly and delicately to avoid physical injury. Parts must frequently be grasped at a specified point.
e. Avoiding chemical contamination of parts. Certain parts, and certain areas of other parts may not be touched with the fingers.
f. The proper positioning of parts between the two electrodes.

It will be noted that the last four responsibilities serve to define the degree of control which the operator must exercise over the motions of her hands and fingers. Since the abilities involved in the rapid, repetitive positioning of parts between the two electrodes play a prominent part in determining the competence of the operator, further description is appropriate at this point.

There are several factors which serve to set the limits of precision within which the operator must position the parts to be welded. Most obvious, perhaps, are the tolerances designated in the engineering specifications. These describe the allowable deviation from the specified location of the parts relative to each other. Tolerances of one-half and one millimeter are common in the radio tube industry. Operators do not work directly from engineering specifications. Instead, throughout the training period, the supervisor endeavors to narrow down the range of the trainee's variations until they fall within acceptable limits.

The available flat surfaces at the end of the electrodes which are in contact with the parts to be welded (Figure 1) limit the operator's freedom in positioning the parts, in a number of ways. Since the two available flat areas are not generally of the same size, the operator must position the parts so that the smaller surface will be located properly on the work at the moment of welding. In the school the movable electrode consisted of a cylindrical welding rod, filed to a taper at one end, leaving an available flat contact surface approximately 0.07" in diameter. The parts had to be positioned on the bottom electrode in such a way that this surface at the end of the movable electrode would bear properly on the parts when the right pedal was depressed.

There are two additional factors which complicate the proper positioning of the parts between the electrodes. Since the movable electrode (Figure 1) is held by the swinging arm (D) which rotates about the shaft (M), the path of the flat contact surface at the end of the electrode is not a straight line. However, when the movable electrode begins to exert pressure on the parts to be welded, its contact surface must be parallel to the contact surface of the stationary electrode, to avoid indentation or burning of the parts. Moreover, from the moment the movable electrode touches the work, the arc-segment which describes its path toward the stationary electrode must be approximately a straight line, perpendicular to the stationary contact surface. If this were not the case, there would be slippage between the two contact surfaces, causing the parts to be welded to move relative to each other. If this occurred and were resisted by the operator, deformation of the work would result.

Furthermore, although the contact surfaces must remain parallel to each other while in contact with the work, it is not necessary that they lie in horizontal planes, as shown in Figure 1. Frequently, the geometry of the parts to be welded, the need for providing operators with adequate visual cues and the requirements governing the movement of the two contact surfaces relative to each other while bearing on the work, dictate filing the parallel contact surfaces in a plane sloping toward the operator. Both horizontal and sloped contact surfaces were used in the school.

Translating these conditions into geometric requirements which must be met by the operator in positioning the parts between the two contact surfaces, we have the following:

a. The two parts to be welded must lie in planes parallel to the planes of the two contact surfaces.
b. The two parts must be placed relative to each other in such way that the compressive forces exerted by the two parallel contact surfaces do not cause movement of the parts relative to each other.

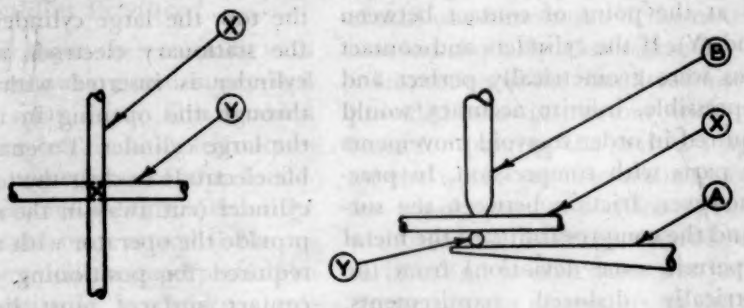
These general statements will be clarified by an examination of the three types of welds made in the Vestibule Training School and common in tube assembly operations.

3. *Three Types of Welds.*

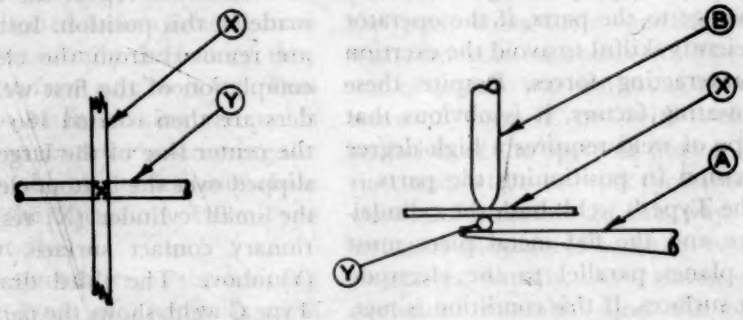
Figure 2 shows the three types of welds made in the school. For convenient reference, these are designated by the letters A, B, and C. In the first sketch of each series, the parts (X) and (Y) are shown welded together, with a small "x" marking the location of each weld. At the right, the parts are drawn properly positioned between the stationary electrode (A) and the movable electrode (B).

There are two geometric requirements which govern the positioning of the two

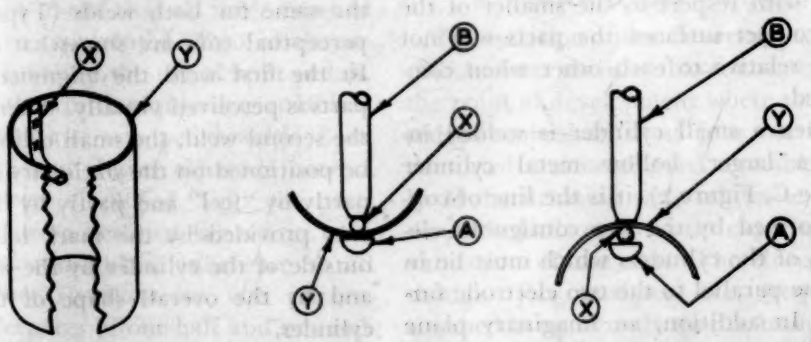
small metal cylinders shown in Type A, Figure 2. First, the cylinders, (X) and (Y), must lie in planes parallel to the planes of the electrode contact surfaces. In addition, the parts must be positioned in such a way that a straight line could be drawn perpendicular to the two contact surfaces at or near their respective centers, which would intersect the two lines of center of the cylinders. If both



TYPE A. TWO CYLINDRICAL WIRES CROSSED AT RIGHT ANGLES.



TYPE B. A FLAT METAL PIECE TO A CYLINDRICAL WIRE.



TYPE C. A SMALL CYLINDRICAL WIRE INSIDE A HOLLOW METAL CYLINDER.

FIG. 2. Three Types of Welds Made in the School.

requirements are met, the imaginary line will also intersect four surface elements of the cylinders; namely, the two elements at the lines of contact between (X) and (B), and (Y) and (A), respectively, and the two elements which cross at right angles at the point of contact between (X) and (Y). If the cylinders and contact surfaces were geometrically perfect and incompressible, infinite accuracy would be required in order to avoid movements of the parts with compression. In practice, however, friction between the surfaces and the compressibility of the metal parts permit some deviations from the geometrically deduced requirements. Moreover, the pressure of the parallel contact surfaces will correct for minor deviations from proper alignment without damage to the parts, if the operator is sufficiently skilful to avoid the exertion of counteracting forces. Despite these compensating factors, it is obvious that this type of weld requires a high degree of precision in positioning the parts.

In the Type B weld, both the cylindrical wire and the flat metal piece must lie in planes parallel to the electrode contact surfaces. If this condition is met, and if the small area in which the parts are to be welded is reasonably well centered with respect to the smaller of the two contact surfaces, the parts will not move relative to each other when compressed.

When a small cylinder is welded inside a larger, hollow metal cylinder (Type C, Figure 2), it is the line of contact formed by the two contiguous elements of the cylinders which must lie in a plane parallel to the two electrode surfaces. In addition, an imaginary plane passed through the lines of centers of the cylinders, must be perpendicular to the electrode contact surfaces, and intersect the smaller one at or near the center.

The sketch for this type of weld shows that two welds are required to join the parts securely. The two methods of positioning the parts are drawn as they would appear to an observer seated in the operator's chair. In making the weld nearest the top, the large cylinder is placed on the stationary electrode and the small cylinder is inserted with the tweezers through the opening in the bottom of the large cylinder. To enable the movable electrode to clear the top of the large cylinder (cut away in the sketch) and to provide the operator with the visual cues required for positioning, the electrode contact surfaces must lie in parallel planes sloped toward the operator (p. 4). The second weld, however, is too far below the top of the cylinder to be made in this position. Instead, the parts are removed from the electrodes upon completion of the first weld. The cylinders are then rotated 180 degrees about the center line of the large cylinder and slipped over the bottom electrode so that the small cylinder (X) rests on the stationary contact surface, with cylinder (Y) above. The third drawing for the Type C weld, shows the parts in this position.

While the geometric requirements are the same for both welds (Type C), the perceptual cues are somewhat different. In the first weld the alignment of the parts is perceived visually. Whereas, with the second weld, the small cylinder must be positioned on the stationary electrode partly by "feel" and partly by the visual cues provided by the mark left on the outside of the cylinder by the first weld, and by the overall shape of the large cylinder.

When one adds to those geometric limitations, the careful handling required by the delicacy of the parts and the need for rigorous adherence to engi-

neering specifications and tolerances, it is obvious that the positioning of the parts for welding demands a high degree of control over the movements of the hands, fingers and tweezers.

4. *The Assemblies Produced.*

All operators began their training with a period of instruction and supervised practice with scrap wire. This enabled them to acquire certain basic techniques and to learn something of the fundamentals of resistance welding without the fear which might have attended the initial use of more delicate good parts. The operation consisted in joining two lengths of 40 mil (.040" diameter), hardened nickel wire, crossed at right angles to their respective midpoints. Since the wires were collected as scrap metal from a trimming operation in the factory, they varied in length from 0.75" to 1.50". Two pedals were used in making the welds and the requirements for positioning the parts between the electrodes were those stipulated in the discussion of the Type A weld (Figure 2).

During the remainder of the training period, three types of assemblies were produced for use in completed vacuum tubes. These will not be described in detail, except to say that all three types of welds were represented. Two of the assemblies involved the use of both pedals, whereas the third required a combination of one and two pedal welds. Both horizontal and sloped electrode contact surfaces were employed. Since the parts produced were destined for use in completed vacuum tubes, adherence to all engineering specifications was imperative. Tolerances of one-half and one millimeter were common. For reference purposes, the three assemblies will be designated by the letters, L, M, and N.

After a period of instruction and supervised practice with the first assembly, the trainee was given a series of *production tests* of one hour duration. At the conclusion of each test, the number of units produced was recorded and the work inspected. When a satisfactory level of skill was demonstrated, the next operation was similarly administered.

Inspection of the completed assemblies provided the instructors with several bases for judging the trainee's aptitude for the work: (a) The trainee's capacity for rapid, precise positioning of parts with respect to both the electrode contact surfaces and to each other was judged by (i) the amount of production during the one hour production tests, (ii) burning indentation or other deformation of the parts, and (iii) the extent of deviations of the parts from their specified location with respect to each other. (b) The proper use of pedals was indicated by the absence of "burned" or "cold" welds. Observation of the operator during instruction and while at work provided additional information. In this way, the instructors endeavored to attain two major objectives, namely, the training of inexperienced mounters and the measurement of their aptitude for the job.

B. THE APTITUDES REQUIRED

The isolation and measurement of specific human abilities have not reached the point of development where the aptitudes involved in these jobs can be inventoried and catalogued unambiguously. If the situation were otherwise, there would be little need for the experimental validation of the tests. Therefore, even though job analyses are frequently couched in apparently specific terms such as "finger dexterity," "foot-eye-hand coordination," and the like, those terms and statements involving them must be regarded for the most part as descrip-

tive of what the operator does. Highly generalized, phenotypical descriptions have many legitimate uses in the employment office and elsewhere. In order to be operationally meaningful in designating sperific aptitudes, however, it would be necessary that each of the terms refer to a specific set of operations (i.e., to a particular method, technique or test) by which the designated aptitudes could be identified and measured, and that the same set of operations could be used to select applicants for any job in which a high degree of the characteristic were found. For example, it would be necessary either to demonstrate that "finger dexterity" refers to the same aptitude in watchmaking, mount assembly, typing or playing the piano, or to develop subclasses of the concept which could be applied with adequate specificity.

These requirements are only partially fulfilled. For, while such terms do not generally refer to a specific set of operations, they do, in some cases, point to a broad class of operations which have been found by factor analysis and correlation studies to be relatively distinct. To the extent that this is true, they serve to delimit the search for specific tests and techniques which are likely to correlate satisfactorily with job performance.

With these limitations in mind, the job was analyzed in accordance with the schedule developed by the War Manpower Commission (24), which rates each of 47 "worker characteristics" on a four point scale. Characteristics receiving either of the two highest ratings, A or B, are listed below.[3]

Working rapidly for long periods	(B)
Dexterity of the fingers	(B)
Dexterity of hands and arms	(B)
Eye-hand coordination	(A)
Foot-eye-hand coordination	(B)
Coordination of independent movements of both hands	(A)
Estimate size of objects	(B)
Perceive form of objects	(B)
Keenness of vision	(A)
Muscular discrimination	(B)
Estimating quality of objects	(B)

The present experiment is concerned primarily with the validation of tests related to dexterity of the fingers, hands and arms, eye-hand coordination and the coordination of independent movements of both hands.

C. PRODUCTION DATA EXAMINED AS POSSIBLE CRITERIA

The possibility of using the production test scores (p. 9) as an independent measure of the ability of each trainee to perform these delicate operations was fully explored.

The reliability coefficients for scores, based upon one, two, three and four successive production tests on each assembly are shown in Table 1. The italicized figures represent the calculated product moment correlations. For example, the index for two trials on assembly L (.893) was determined by correlating the total number of units produced during the first two production tests with a similar score for the third and fourth tests. The italicized coefficient for as-

TABLE 1

Reliability of Scores Based Upon One, Two, Three and Four Successive Production Tests on Each Assembly

Operations	N	Number of Production Tests			
		1	2	3	4
Assembly-L	203	.807	*.893*	.926	.943
Assembly-M	200	.702	*.825*	.876	.904
Assembly-N	200	*.727*	*.842*	.888	.914

[3] The author is indebted to Mr. L. A. Kameen, experienced job analyst, for checking the analysis for technical accuracy.

sembly M (.825) was computed in the same way. Intercorrelation of the number of units fabricated during the first and second tests on assembly N, yielded a coefficient of .727, the reliability of the scores on one production test. The remaining indices were computed with the Spearman-Brown prophecy formula. These data were obtained by selecting at random approximately two hundred records from the many hundreds who were trained in the school during the first ten months of the year, 1943. This was done in order that the results would be applicable to all the experimental data collected during that period. Since not all trainees had the required number of production tests on each operation, the sampling process was repeated for each coefficient.

The results indicate that the production tests have attained a degree of reliability comparable to that of many employment tests. One or two trials on each assembly would provide a score sufficiently reliable for mass testing purposes. There were, however, several factors which limited the value of production data as criteria.

1. Unequal amounts of practice before beginning the first production test on a job. Although a given group of trainees, entering the school on the same day, generally followed the same schedule, there were variations among the different groups processed over a period. Moreover, trainees whose performance failed to meet certain minimum requirements were frequently given additional practice before taking the first test.

2. Variations in the number of production tests given on each job. Production tests were continued with each assembly until the trainee achieved a satisfactory level of production and quality. Thus the better operators generally had less experience in the first welding operation before undertaking the second and third assemblies. The number of production tests administered also depended upon the relative number of each type of assembly required by the factory over a period of time as well as upon the availability of materials.

3. It was not always possible to adhere to the planned sequence of jobs. For example, it was sometimes necessary to begin a trainee on assembly M without the benefit of experience with assembly L.

4. Variations in quality and workmanship. Since speed of production depends to a least some extent upon the amount of attention given to quality and workmanship, raw production test scores are not, in themselves, precise measures of differential ability.

5. Although every effort was made to avoid interruptions during production tests, it was frequently necessary to correct errors in performance as they appeared, in order to prevent their fixation. Also, occasional machine and material difficulties are inevitable over an experimental period of several months.

To the extent that the first three factors operated, trainees were tested at different points on the learning curve. While these variations may be expected to have little effect upon the reliability with which the trainee's skill at the time of testing is measured, they would tend to reduce the intercorrelations among the scores for the three assemblies, as well as their suitability as criteria.

Variations in quality, on the other hand, would reduce the reliability coefficient for any test if they occurred from trial to trial for the same individual. Differences in quality from job to job for the same individual would reduce interassembly correlations. Any variance in quality among individuals would reduce the validity of the scores as criteria.

The occurrence of interruptions would tend to reduce all correlations from their true values.

While one might wish that these sources of variation had been controlled, the fact that the techniques and procedures themselves have been demonstrated to be reliable measures of the abilities required to perform each of these jobs, is

a strong argument for the validity of the instructors' judgments based upon these scores. If the techniques had been found to be unreliable, the validity of their judgments would be open to serious question. Given reliable techniques, however, it was possible for the instructors to correct for variations which attended the testing of individual trainees. To have accomplished this with a high degree of precision, it would have been necessary to develop a set of norms for each particular set of circumstances. Since multiple norms were not available, it would be naive to pretend that the corrections were applied without error. However, the instructors had had opportunity to observe and judge many hundreds of trainees under similar conditions, before the present experiment was begun. It is reasonable to assume, therefore, that their pooled judgment, based upon several days of constant observation of trainee performance, would constitute a better criterion than the uncorrected production test scores themselves.

The criterion finally selected will be described in the following chapter.

III. The Correlated Variables

The validation of aptitude tests consists essentially in determining the degree of relationship between two sets of variables, namely, the criterion of job performance and the several aptitude test scores.

A. THE CRITERION

The criterion was a single overall rating, based upon the pooled judgment of the supervisor of the school and at least two instructors, expressed on a five-point scale. The procedures in the school were such that each instructor had the opportunity to teach every trainee at least one of the assemblies and to administer the required production tests. The supervisor of the school followed closely the progress of each trainee by repeated observations of performance, by assisting the instructors in teaching the operations and by checking the reported scores on the production tests as well as miscellaneous notations on the record sheet. In addition, an effort was made to size up the trainee's probable adjustment to and attitude toward the work, by means of many informal contacts and one planned interview.

Upon completion of the training, the supervisor, together with the instructors, reviewed the trainee's performance and general fitness for the occuption. The overall rating was arrived at through discussion and represented the consensus of the group.

1. *The Meaning of The Overall Rating Continuum.*

Since the factors which make up this composite rating determine entirely the usefulness of the correlations obtained with the tests, the meaning of the continuum will be explored.

(a) *Verbal definitions.* Two separate written definitions were solicited from the supervisor of the school, the first during the course of a preliminary experiment and the second more than six months later, immediately before the present study. Both statements agree that three factors must be considered in judging a trainee's probable success as a mounter: (i) *ability to learn,* indicated by the speed with which instructions were grasped and the extent to which they were retained; (ii) *dexterity,* judged by ability shown in handling tweezers and small parts and by the speed and precision demonstrated in performance of the required operations; and (iii) *attitude,* based upon considerations such as willingness to work, industry, patience, reaction to criticisms, attitude toward quality and conscientiousness.

Of course tests of manipulative aptitudes can hardly be expected to predict the types of behavior which were included in the concept of "attitude." Nor were they designed to measure the ability to follow and retain verbal instructions. A criterion excluding these factors, however, would fail to provide an index of the importance of the aptitudes measured by the tests in the total contellation of factors which make for job success. For this reason, the overall ratings were preferred to judgments based on manipulative abilities alone.

(b) *Statistically determined definitions.* Two statistical studies were made which served to determine at least partially the meaning or content of the overall ratings. The first was a byproduct of an earlier experiment and shows the relationship between judgments of quantity and quality and an overall evaluation. In the second, the ratings are correlated with raw

production test scores for each of the three assemblies.

(i) In order to obtain criteria for the first 51 trainees tested in the series of experiments involving the school, the ranking method was employed.[4] When the 51st employee had completed the training, a set of cards bearing only the names of the employees was ranked according to overall fitness for the occupation. One week later, two new sets of 51 cards each, for the same employees, were ranked, respectively, for the quality and for the quantity of the work produced during training. The rank-difference correlations (ρ) between the three sets of ranks are shown in Table 2. The denominator of the corresponding t ratios was calculated by setting ρ equal to zero in the formula for the standard error of that statistic. At the 1% level, with 49 degrees of freedom, t equals 2.68. Thus all intercorrelations are highly significant.

TABLE 2

Intercorrelations Between Quality, Quantity and Overall Ranks
(N = 51)

Ranks Correlated	ρ	t_ρ
Quality-Quantity	.665	4.54
Quality-Overall	.740	5.05
Quantity-Overall	.852	5.81

The correlation of .665 between the quality and quantity ranks indicates a substantial amount of overlap between the judgments of these two factors. Granting validity to the respective judgments, this would mean that the faster operators tend to produce work of higher quality, whereas the slower operators tend to produce work of poorer quality.

The remaining two coefficients indicate that the separate judgments of quality and quantity are important components of the overall judgment. The somewhat greater weight given to quantity of work may be explained in part, at least, by the circumstances attending the ranking. The 51 subjects had completed their training during the three week period immediately preceding the overall ranking. An interval of between one and four weeks elapsed before the quality and quantity rankings were made. It was necessary, therefore, for the supervisors of the school and the instructors to depend upon their written records of performance. Quantity of production on each of the jobs, together with notations referring to extraneous factors (e.g., machine trouble) affecting it, were recorded more precisely than characteristics such as workmanship and attitude. It would not be surprising, therefore, if quantity of work was given more weight in these belated rankings than it would have received if the judgments were made immediately after graduation.

These data are important to the present experiment in several ways. They demonstrate that the pooled judgments of the supervisor and the instructors are capable of high reliability, and, that quantity and quality of work are important components of the overall judgments. Since the records were consulted throughout, the results lend support to the assertion that reliable judgments can be made on the basis of the production test scores, the number of defective units produced and notations referring to workmanship, attitude and disturbing conditions.[5]

[4] The rank method was preferred to rating in this preliminary experiment because it was thought to involve a simpler and more precise judgment process. The results were employed, along with discussions of the nature and distribution of individual differences, as a means of training the supervisors in making the ratings which served as a criterion for evaluating all tests given subsequently to prospective mounters.

[5] It cannot be stated, however, that the judgments were based *solely* on the records. Since the

(ii) The extent to which the overall ratings are based upon the magnitude of the production test scores is indicated by the correlation coefficients in Table 3. The raw data were selected from the files in the same way, and for the same period, as those used in computing the reliability of the respective assemblies. The raw scores on each assembly were comprised of the total number of units produced during the first two production tests.

TABLE 3

Correlations Between Overall Ratings and Raw Production Test Scores

Operation	N	r_{bis}	$\sigma_{r_{bis}=0}$	t
Assembly-L	192	.646	.0987	6.55
Assembly-M	208	.603	.0904	6.67
Assembly-N	200	.525	.0928	5.66

For reasons to be discussed in Chapter V, the rating distributions were divided dichotomously and the coefficients were calculated biserially. The denominator of the t ratio was computed by setting r_{bis} equal to zero in the formula for the standard error of the biserial r (14). Since t values of 2.61 and 2.60 are significant at the 1% level for 150 and 200 degrees of freedom, respectively, all coefficients are highly significant.

It was noted earlier that the supervisors endeavored to correct the raw scores for the extraneous factors affecting them. Having demonstrated the reliability of the supervisors' judgments, based on production and other records, the coefficients may be interpreted as representing the minimum correlation between quantity of production on each of the jobs and the overall ratings. Ordinarily, the multiple correlation coefficient would be calculated to determine the extent to which the overall ratings were based upon a composite of these scores. However, since all zero-order coefficients were distorted by uncontrolled conditions (pp. 11f) this statistic could not be taken as a good estimate of the amount of relationship. There is little doubt that the true multiple R, if known, would be greater than the highest zero-order coefficient in Table 3.

In spite of this, the data further substantiates one of the conclusions reached earlier, namely, that quantity of production is an important component of the overall ratings.

2. *The Steps of The Rating Scale.*

The rating scale consisted of the five grades, Poor, Fair, Average, Good and Very Good, which were selected by the supervisors during the discussions in which the procedures were developed. The overall rating received by a trainee was communicated to the experimenter as an abbreviation of the appropriate adjective, recorded on a card bearing the employee's name, the number of units produced during each production test, and miscellaneous observations and comments. While each of the adjectives was verbally defined, their essential meaning for this experiment will be derived (Ch. V, Section A) from an analysis of the distribution and from the answer to a series of questions, designed to separate "satisfactory" from "unsatisfactory" employees.

B. THE APTITUDE TESTS

Five manipulation tests were administered to all applicants included in this

individual trainees were once known to the instructors, it is probable that memory of their performance played a part in determining both sets of rankings. However, with respect to the reliability of the rankings, it is necessary only that the spurious effect of memory of the first ranking be eliminated. It appears that this factor was adequately controlled in the experimental design, by the number of subjects, the time intervals, and the fact that the judges were led to believe that the overall rankings completed the task.

study. Although a standard intelligence test was also given when time permitted, a check on the mean and spread of the scores showed that it was not discriminating sufficiently among the lower scoring members of the applicant group. Since the selection and validation of a more suitable measure of intelligence did not coincide with the defined boundaries of the present experiment, these data will not be presented.

1. *Procedures for Administration.*

The procedures for administration and scoring were adapted to the requirements of the mass testing situation in which the results were to be applied. The revised procedures are described below. The tests were administered in the order listed.

(a) *Minnesota Rate of Manipulation, Placing.* The standard instructions (17) advise the subject to "Push the board away from you, leaving a clear space about twelve inches wide in front of you." In order to control the distance of the hand movements and the alignment of the blocks in relation to the empty board, four wooden angles were secured to the table. In this way, a constant distance of ten-and-five-eighth inches was substituted for the somewhat variable twelve inch distance recommended in the instructions. The basic motion pattern remained the same.[6]

In contrast with the standard instructions, the test was administered on a time limit basis. After one complete trial for practice, four forty-second trials were given. The score consisted of the total number of blocks placed during the four trials. After each trial, the subjects were instructed to place the remaining blocks in the holes, using both hands. The starting directions, which are important motivationally, were as follows: "Now you want to see how many you can do in just 40 seconds. Take hold of the bottom block with your right hand. Get ready. Go." Before the second trial, the need for repetition was explained as a means of giving everyone a fair chance. This was found necessary in order to maintain cooperation throughout and to avoid the "What, again?" attitude which had been expressed frequently.

(b) *Minnesota Rate of Manipulation, Turning.* Except for the substitution of the time limit for the work limit method, this test was administered in the standard way (19). After one complete trial for practice, four thirty-second trials were given. The total number of blocks turned during the four trials comprised the score, After each trial, the subjects were instructed to complete the board exactly as during the test. The send-off directions were as follows: "Now you want to see how many blocks you can turn over in just 30 seconds. Take hold of the first block in the upper right hand corner with your left hand. Get ready. Go."

(c) *Finger Dexterity Test.* The test equipment, described by O'Connor (11) was used. According to the standard instructions (21), subjects are directed to "Start in the farthest corner and work toward you." On the other hand, instructions to the examiner say: "Allow the examinee to place 30 pins, thus filling the top line of ten holes, for practice." Apparently following this latter instruction, Tiffin (18) presents a picture of a subject filling the holes from left to right along the top of the board. At the same time, O'Connor (11) shows a subject taking the Tweezer Dexterity Test, for which the directions are the same in this respect, who has filled five rows perpendicular to the length of the board and who appears, moreover, to be placing the first pin in the sixth row in the hole nearest her. Thus it is difficult to say whether or not the standard procedure was followed in this experiment. Nevertheless, all applicants were instructed to start in the upper left hand corner and to work down the row, that is, transverse to the length of the board. All

[6] This arrangement was found satisfactory for the period in which the placing and turning tests were administered experimentally to all applicants, and when space permitted the permanent storage of the tests on the tables. It was modified, however, when these tests were found valid for only a few occupations. According to the new method, the tests are stored in a cabinet on removable, plywood shelves, with a 5/16" high, square, wooden strip along each of the four edges. The strips along the two side edges are cut away over a length of four inches to permit lifting the empty test board after it is used to position the blocks for placing. All dimensions are such as to preserve the original conditions in which the test was standardized.

rows were filled in the same way. Since the time limit method was used and only a portion of the board completed, the order in which the holes are filled is important because of its effect upon the total distance travelled in the testing period.

After filling the ten holes in one row for practice, the subject was given two three-and-one-half minute trials, starting each with an empty board. The score consisted of the total number of holes filled during the two trials. The time allowed was stated in the starting directions for this as for all tests.

(d) *Purdue Pegboard,*[1] *Assembly Test.* The subtest of the Purdue Pegboard (22) included in this study, consists of the repetitive assembly of a pin, a cylindrical collar and two washers, and will be referred to as the Purdue Assembly test. Except for the fact that four complete assemblies were allowed for practice in addition to the two normally constructed in the course of the initial instructions, the standard procedures were followed. The test proper consisted of three separate trials of one minute duration. The number of parts assembled during each trial was recorded and their total served as the final score.

[1] Tiffin (18) refers to this test as the "Purdue Dexterity Test."

(e) *Tweezer Dexterity Test.* The standard test equipment, designed by O'Connor (11, 23), was used. Holes were filled in the order described for the Finger Dexterity Test. In contrast to the method pictured by O'Connor (11), the tweezers were held somewhat like a pencil, with the tip of the third finger near the point. The index and third fingers were spread apart on one leg of the tweezers, with the thumb pressing against the opposite leg. A similar technique is applied to most mounting operations. The method of placing the pins was also covered in the instructions: ". . . if you pick up the pins correctly and keep your hand and wrist relaxed, you can slant the pin to start it in the hole and then straighten it out. . . . But be sure to straighten it out before you let go of it, otherwise it stands up like this and that's wrong."

After filling the ten holes in the first row for practice, the subject was given two three-minute trials, starting each with an empty board. The score consisted of the total number of holes filled during the two trials.

Data related to the reliability of the tests as applied to the experimental group will be presented in Chapter V.

IV. Experimental Conditions and Subjects

Except for the administration of aptitude tests, all subjects were processed in accordance with the regular selection and placement of procedures in effect at the time. As applied to female applicants for production jobs in the factory, these procedures were essentially as follows:

1. All applicants were required to fill in an *application blank*.
2. They were then given a *screening interview* for the purpose of checking such details as completeness of the application blank, type of work desired, and shift preferred. An effort was made to determine in a general way whether or not some type of available factory work was mutually acceptable.
3. The *experimental battery of tests* was administered in a quiet, pleasant room somewhat removed from the activities of the employment office. All applicants took the tests under the impression that the results were to be used for placement purposes. Actually, the results were not reported to the interviewers. Instead, they were filed and later compared with performance records.
4. During the *placement interview* which followed, each member of this applicant group was assigned to a particular occupation and applied against a specific requisition.
5. Applicants were then routed to the dispensary for a *medical examination* which included such items as general physical condition, height and weight measurements, and a telebinocular test of vision.

Individuals designated as mounters were scheduled to begin training on the following work day in the Vestibule Training School. After three or four days of training, they were assigned to specific operations in the factory, in accordance with the abilities and characteristics demonstrated in the school.

It was noted earlier that, since the research with mounters covered a considerable period of time and was pursued with various tests and procedures, the present study does not include all applicants processed in this way. The major reduction in the size of the group was accomplished by fixing the dates for the beginning and end of the experiment, as May 17th and September 15, 1943, respectively, during which time the tests and the procedures for administration followed rigorously the pattern previously described.

Application of the following criteria resulted in the further elimination of subjects from this group.

1. All subjects known to be left handed were eliminated to avoid the possible spurious effect of this factor on the empirical correlations. Left handedness was determined in two ways. (a) All subjects were allowed to choose the hand to be used for the placing test. In the event of hesitancy, the importance of using the faster hand was emphasized, and further opportunities for making the choice were presented during the instructions on the finger dexterity and tweezer dexterity tests. A question mark was posted to separate those who vascillated or changed hands from those who were consistently right handed. (b) The supervisor of the school made frequent notations on the cards reporting production test scores and overall ratings, in order to assist the experimenter in identifying factors which might affect the relationships between tests and performance. These notations were also posted on the trainee's card. Although the method of determining handedness in individual cases was not recorded, all records bearing the notation, "L.H.," with or without a question mark, were excluded.
2. In order to avoid recalculating certain intermediate statistics for each correlation coefficient, subjects who did not have all five manipulation tests were eliminated.
3. Individuals who had previous experience with the company were not included. Since this information was posted from the daily hire sheet which identifies rehires without specifying the previous occupation, it is possible that a number of the trainees excluded had no direct experience as mounters. However, the fact that most tube making

operations provide experience in the manipulation of small parts and, in some cases, training in the use of tweezers, the elimination of all rehires was considered advisable.

4. An effort was made to detect identical or related experience in other companies. This was done in two ways. (a) Subjects were asked, during the course of the instructions on the tweezer dexterity test, "Has anyone used tweezers in work before?" (b) Previous experience in mounting was one of the factors noted on the report card from the Vestibule Training School, and was posted to the trainee's file card. Since all trainees were thoroughly interviewed in the school, it is reasonably certain that all experienced mounters were detected and eliminated. However, since applicants were drawn from a highly industrialized area, it is entirely possible that all experience helpful to performance in both the test and job situations was not adequately controlled. Moreover, while the point would be difficult to check at this late date, it is the author's impression that no one answered affirmatively the question pertaining to the previous use of tweezers who would not have been eliminated as an experienced mounter. If this is so, it raises the question of whether, in view of the motives for concealing this information in the test situation, the answers were valid. Further data, pertinent to this question as it relates to the tests of the final battery, will therefore be presented in a later section (Ch. V, Section I).

5. Only those subjects who were reported to have had a total of three or more production tests on at least two of the three standard jobs in the training school were included. The fact that many of the eliminated subjects had been given tests on other jobs was not considered. While there were good reasons for requiring production tests on all three jobs, this stricter criterion would have been selective with respect to the abilities involved. Poor and very poor trainees were given considerably more training and practice on the first two jobs than is usually required, and were frequently assigned to carefully selected non-mounting jobs in the factory before completing the training. The experimental group, which consisted of 233 trainees, received a total of 1895 tests on welding jobs while in the school. 1791 of these were distributed over the three welding jobs described, yielding an average of 7.7 tests per trainee for these three jobs. Considering the reliability coefficients presented in Table 1, there can be little question that the related abilities of the majority of the experimental group were adequately measured in the school. If the group includes individuals who were judged on unreliable or different grounds, these would serve to reduce the correlations between the tests and the criterion. Thus, if satisfactory validity coefficients are obtained, it can be assumed that subjects were judged with sufficient reliability and uniformity.

6. Trainees who were rated by the instructors during the period when the supervisor of the school was on vacation were eliminated.

In this way, a group of 233 subjects was selected for further study.

V. Results and Discussion

Two independent sets of data were thus procured in the course of the experiment, namely, the ratings made by the supervisor of the school and the staff of instructors, and the respective test scores. It remains to present and examine both sets of data, to determine the degree of relationship between each of the tests and the criterion and to develop a regression equation which will predict performance in the school with a minimum of error.

A. The Criterion Scores

The number and percent of subjects receiving each of the five overall performance ratings are presented in Table 4. Inspection of the data reveals an excess of ratings at the high end. The extent of these deviations from normality and the likelihood that they could have arisen by chance alone will be determined by applying the tests for skewness and kurtosis to the data.

Table 4
Distribution of Overall Ratings
(N = 233)

Rating	N	%
Poor	21	9.0
Fair	36	15.4
Average	67	28.8
Good	57	24.5
Very Good	52	22.3

1. Analysis of Distribution for Skewness and Kurtosis.

Assigning the numbers, 1, 2, 3, 4, and 5 to the steps of the rating scale, starting with Poor, the mean of the distribution becomes 3.86, where 3.50 is the theoretical midpoint of the scale. The standard deviation, in terms of the assigned numbers, is 1.24. The test for skewness yields a value of −0.278 for g_1 (15), with a standard deviation of 0.159 and a t of 1.75. Since the test of significance is carried out with an infinite number of degrees of freedom, t must equal or exceed 1.64, 1.96, or 2.33, in order to be significant at the 10%, 5% or 1% levels, respectively. Thus, while not significant at the 5% level as usually defined, a negative skewness as great or greater could arise by chance, under these conditions, only about 4.5% of the time (14). Further evidence is found in the distribution of 359 ratings, obtained under similar circumstances, immediately prior to the present experiment. The number and percent of subjects receiving each of the five ratings from Poor to Very Good were 32 (8.9%), 59 (16.4%), 86 (24.0%), 96 (26.7%) and 86 (24.0%). It appears safe to conclude, therefore, that the negative skewness of the present distribution did not arise by chance alone.

The statistic, g_2, was calculated to be −0.854, with a standard error of 0.318. Since t, for these values, is 2.69, platykurtosis is demonstrated at better than the 1% level.

Thus the criterion distribution departs significantly from normal with respect to kurtosis and exhibits a strong tendency toward negative skewness.

2. Differential Reliability Along The Rating Continuum.

The activity of judging performance in terms of the steps of a rating scale must be viewed against the backdrop of the needs and the practical demands of the situation, and the rater's reaction to them, if the results are to be interpreted and evaluated properly.

One aspect of the situation which undoubtedly affected the ratings was the fact that the supervisor of the school had to decide among three possible courses

of action for each employee. (a) *Termination.* Since the trainees were already employees of the company, every effort was made to avoid this action. (b) *Placement on a relatively simple job other than mounting.* Because the demand for mounters was greater than for any other single occupation, considerably more time was spent in evaluating the aptitudes of the poorer operators. This was necessary to avoid excluding trainess with a reasonable chance of success in mounting and to assess other potentialities and interests which might indicate placement on another type of work. (c) *Placement in the factory as a mounter.* This course of action frequently entailed additional decisions relating to the difficulty of the job which could be handled by the trainee. In addition, while the responsibility for final placement rested with the school supervisor, the number of operators to be assigned to any specific manufacturing area during any period was controlled by a priority system, developed as a wartime measure and operated by those in charge of planning and scheduling production. Thus, even though considerable freedom was allowed in order to insure maximum long-range utilization of the manipulative and other abilities demonstrated in the school, the practical question frequently became, "Can this trainee do the job to which we should assign her?" Under these circumstances, it is not surprising that more time was spent determining whether or not the trainee had the minimum degree of manipulative ability required by the specific job under consideration, than in carefully measuring the differences in ability among those whose success seemed assured.

Of equal importance is the fact that the poorer operators required more training and attention than those who grasped instructions quickly and encountered few problems. It is only natural that the instructors should have allocated their time in accordance with the needs of the individual trainees. The only exception to this general tendency occurred in the handling of trainees assigned to mounting operations involved in exceptionally small and delicate types of tubes. These girls were selected from the Good and Very Good groups during their last day in the school and were given training, practice and production tests with a finer, more difficult type of assembly. But these additional measures did not result in a more precise grading of all the trainees in the two highest rating groups. Not only were many trainees with high ratings sent to other jobs, but the selection of trainees for these operations depended upon the requirements of the factory at the particular time.

Thus, because of the practical demands of the situation in which the judgments were made, it is highly probable that employees at the lower end of the scale were rated with greater precision than those at the higher end. These facts had an important bearing upon the selection of statistical methods for the remainder of the study.

3. *The Selection of Appropriate Statistical Methods.*

The existence of differential reliability between the higher and lower steps of the scale points to the biserial r as the statistic which will properly evaluate the relationships between the test scores and the criterion. According to this method, the criterion distribution is divided into two parts by drawing a line through a suitable point, somewhere along the rating scale. Trainees with higher ratings are placed in one group and those rated below the line are placed in a second

group. Since this dividing line could be drawn between any two contiguous steps on the scale, a choice must be made from among four possible lines. Ordinarily, the differences in size among the coefficients yielded by the various divisions would fall within the limits of the standard error of the coefficients; that is, variations in size would be due to chance errors alone. Normally, therefore, the choice of dividing line is of little consequence, excepting in so far as the relative sizes of the two groups affects the magnitude of the standard error itself. However, since the practical situation demanded more precise classification of the group at some points along the scale than at others, variations among the four possible correlation coefficients for any test would probably exceed that assignable to chance errors alone. In order to insure that the choice would be made on grounds other than the characteristics of the resultant arrays, the following interoffice communication was sent to the supervisor of the school.

Taking it for granted that people rated Very Good and Good are good risks for mounting, which of the following statements is most nearly true? Check one.

( ) People rated Poor are not good risks for mounting; people rated Fair and above are good risks.

( ) People rated Poor and Fair are not good risks for mounting; people rated Average and above are good risks.

( ) People rated Poor, Fair and Average are not good risks for mounting; people rated Good and above are good risks.

The sheet was returned with the second statement checked, indicating that trainees rated Poor and Fair were regarded as "unsatisfactory," whereas those rated Average and above are considered "satisfactory." Some time later, while discussing this response, the supervisor remarked that if it were at all possible, she would not place in mounting anyone who failed to achieve the performance standards required for a rating of at least Average.

For these reasons, the dividing line was located between the ratings, Fair and Average, in calculating the biserial correlations between the ratings and the respective tests. Since the biserial r is mathematically equivalent to a product moment r, corrected for broad categories, this procedure provides a statistic which can be used in developing a regression equation. The relatively slight loss in precision which attends the broad grouping of applicants will be reflected in the standard error employed for testing significance.[8]

Another problem posed by the nature of the criterion distribution concerns the scale to which the regression equation must predict. In order to avoid the complex mathematical procedures of curvilinear regression, which predictions to the original platykurtic and negatively skewed distribution would entail, an arbitrary scale was constructed by selecting a mean of 50.000 and a standard deviation of 14.286 (= 100/7), as parameters of the criterion distribution. Thus the range, $M \pm 3.5\sigma$, spans the numbers from zero to one hundred and steps of 0.7σ

[8] In a previous experiment with the same criterion, correlation coefficients were calculated by the product-moment method and biserially, with the dividing line drawn between the Fair and Average ratings. The biserial coefficients were consistently higher. This was interpreted as substantiating the conclusions reached through analysis of the psychological situation of the raters. Strictly speaking, the biserial coefficients represent the amount of correlation that would be obtained by the product-moment method utilizing the five rating grades and correcting for broad categories, if discriminations were made with the same precision between each of the contiguous pairs of ratings on the scale as between the ratings, Fair and Average.

become steps of 10 points on the assumed criterion scale.

The procedures followed to relate predictions on the assumed scale to the five grades of the original rating scale will be described in Section E of this chapter.

B. THE TEST SCORES

1. *Means and Standard Deviations of Test Scores.*

In Table 5 the mean and standard deviation of the scores on each test are entered, together with their respective standard deviations. The means (M) represent the average performance of the experimental group on the tests. The

TABLE 5

Means and Standard Deviations of Test Scores

N=233

Tests	M	σ_M	$\sigma_{dist.}$	σ_σ
Placing	171.9	0.84	12.8	0.59
Turning	156.7	0.95	14.4	0.67
Finger Dexterity	92.0	0.64	9.7	0.45
Purdue Assembly	127.9	1.12	17.1	0.79
Tweezer Dexterity	99.4	0.87	13.2	0.61

precision with which the mean was determined (i.e., the amount of variation among sample means to be expected upon replication) is indicated by the respective values of σ_M. The standard deviation of the distribution (σ dist.) for each test serves as a measure of the extent to which individuals differ in test performance, whereas the standard deviation of this statistic ($\sigma\sigma$) reflects the precision with which it has been determined. It is apparent from these data that the mean and the standard deviation of the distribution have been determined with reasonable precision.

2. *Reliability of The Test Scores.*

Data pertinent to the reliability of the total scores on each test, as applied to the experimental group, are recorded in Table 6. For the placing and turning tests, the initial product-moment r was calculated between the sum of the scores on the first two trials and the sum of the scores on the third and fourth.

The reliability of the scores based upon four trials was then computed by the Spearman-Brown prophecy formula. Thus the indices in the fourth column represent the reliability of the total scores on these tests for the experimental group. The initial r's for the remaining three tests were calculated between the scores on the first and second trials, by the product-moment method, and extended by the prophecy formula for two or three trials as indicated.

For mass testing purposes, the coefficient of reliability is significant primarily because of its effect upon the validity coefficient. Not only does it set a ceiling on possible validity (18), but the chance errors introduced as reliability decreases

TABLE 6

The Reliability of Test Scores

Test	No. of Trials	Time/Trial	Coeff. of Reliability	$P.E._{1\infty}$
Placing	4	40 sec.	.885	2.9 blocks
Turning	4	30 sec.	.822	3.7 blocks
Finger Dext.	2	3.5 min.	.869	2.4 holes
Tweezer Dext.	2	3.0 min.	.820	3.8 pins
Purdue Assembly	3	1.0 min.	.906	3.5 parts

attenuate validity coefficients at all levels. Had these data been available prior to the experiment, the number of trials would have been selected to achieve a more uniform level of reliability on all five tests.

Proper interpretation of an individual's score on a test requires an estimate of the probable divergence of the applicant's true score from the one obtained. In so far as the discrepancy is due to chance errors, such as fumbling, getting off to a bad start, the portion of the last motion cycle incomplete at the stop signal, momentary confusion in method, etc., its probable magnitude can be estimated by the probable error of measurement, given in the last column of the table. For example, the chances are about even that an individual's true score on the placing test will fall within the limits defined by the obtained score ± three (2.9) blocks. In slightly more than four-fifths of the cases the individual's true score will lie within the range, the obtained score ± six (5.8) blocks. Thus small differences among applicants on the tests do not necessarily indicate true differences in the capacities measured. Differences corresponding to those itemized in the table may be entirely disregarded in the employment office.

3. *Intertest Correlations.*

The intercorrelations between the tests of the battery are presented in Table 7. The lowest correlation (.369) is more than twice that required for significance at the 1% level with 231 degrees of freedom (15). Since all correlations are positive, it is obvious that applicants who score highly on one test tend to do well on each of the remaining tests. Similarly, there is a tendency for applicants who do poorly on one test to obtain low scores on each of the others. The strength of these tendencies is indicated by the magnitude of the product moment coefficients in the table.

C. THE CORRELATIONS BETWEEN TESTS AND CRITERION

The extent to which scores on the respective tests are related to the criterion is indicated by the magnitude of the biserial coefficients in the second column of Table 8. The standard error employed in the t test was computed by setting r_{bis} equal to zero in the formula for that statistic (14). Since a t of 2.60 is significant at the 1% level for these conditions, all coefficients are highly significant.

TABLE 8

Correlations Between Tests and Criterion
(N = 233)

Tests	r_{bis}	t
Placing	.564	6.56
Turning	.499	5.79
Finger Dexterity	.482	5.60
Purdue Assembly	.636	7.39
Tweezer Dexterity	.586	6.81

TABLE 7

Intertest Correlations
(N = 233)

	T	FD	PA	TD
Placing (P)	.558	.461	.462	.395
Turning (T)		.459	.511	.369
Finger Dexterity (FD)			.418	.510
Purdue Assembly (PA)				.439
Tweezer Dexterity (TD)				

Having established that each of the five tests is significantly related to the criterion, several questions arise. Should all five tests be administered in the selection of future mounter trainees? If not, which of the tests should be included in the final battery? How should the scores on the selected tests be weighted to insure maximum forecasting efficiency? These and related questions will be discussed in the following section.

D. THE COMPOSITION AND YIELD OF SELECTED TEST BATTERIES

Hull (10), in 1928, clearly demonstrated "the radical tendency to diminishing returns as successive tests are added to the battery," the rate depending upon the absolute and relative sizes of the validity coefficients and the intertest correlations. It has also been known for many years that the multiple correlation coefficient is spuriously high, due to the accumulation of all chance errors in the positive direction. Thus, successive tests usually contribute progressively less to the forecasting efficiency of a battery, while each adds its full share of chance errors. The rate of diminishing returns is therefore greater than indicated by the decreasing increments to the multiple R. More recently, Wherry (16) has added to the Doolittle method (14), a technique which indicates the optimum order in which tests should be added to the battery, starting with the test of highest validity, and which yields a series of "shrunken multiple correlations," $(\overline{R})$, corrected for the cumulative, positive, chance errors added by each successive test.

The results in Table 9 were obtained by applying the Wherry-Doolittle Test Selection Method to the present data. Most striking is the amount of agreement between the regular and the "shrunken" multiple correlation coefficients. Differences are reflected only in the third significant figures. Moreover, the calculated shrinkage decreases from 0.002 to 0.001 with the addition of the fourth and fifth tests, respectively. There are several reasons for this result.

TABLE 9

The Effect of Successive Additions to the Test Battery on its Relationship With the Criterion*

Optimum order of test addition	R	$\overline{R}$	100 $(\overline{R})^2$
Purdue Assembly (Zero-order r)	.636	.636	40.4%
Tweezer Dexterity	.722	.720	51.8%
Placing	.757	.755	57.0%
Finger Dexterity	.761	.760	57.8%
Turning	.762	.761	57.9%

* Optimum order of test addition as determined by Wherry Shrinkage Formula.
R—Multiple correlation coefficient.
$\overline{R}$—Shrunken multiple correlation coefficient.
100 $(\overline{R})^2$—Coefficient of determination, expressed as a percent.

1. The Wherry shrinkage formula contains the fraction, $(N - 1)/(N - M)$, where N is the number of cases and M, the number of variables in the regression equation. Thus, little shrinkage is contributed by this factor when the number of cases in the study is large and the number of variables under consideration is small. In the present study, this fraction differs from one only in the fourth significant figure as the second and third tests are added, and in the third significant figure with the addition of the fourth and fifth tests.

2. The magnitude of the shrinkage is determined partly by the amount of covariance added by a given test. In the case of the fourth and fifth tests, the amount of added covariance is small, as indicated by the increments to the unshrunken R.

3. As with the zero-order r's, Wherry-Doolittle calculations were carried out to four significant figures. Apparently the cumulative effect of rounding the fourth significant figure was sufficient to cause the difference between the regular and the shrunken coefficients to be smaller (by .001) with four

or five tests in the battery than with two or three tests.

The proper testing of the significance of the tabulated R's is complicated by the fact that the zero-order r's were calculated biserially. Although the biserial r is mathematically equivalent to a product moment r corrected for broad categories, its standard error is somewhat larger than when computed by the product moment method. Consequently one would expect that a given multiple R would be significant at a higher level when compounded of product moment r's than is the case when they are calculated biserially. However, since the tabulated R's exceed so considerably the usual requirements for significance at the 1% level (15), a test of greater refinement is rendered unnecessary for these data.

The entries in the fourth column of the table show, for each R, the coefficient of determination expressed as a percent. Since equal increments to this coefficient have the same significance at all points along the scale, it provides a more suitable comparative measure of the contribution of successive tests than either the regular or the shrunken multiple correlation coefficients. Thus, 40.4% of the criterion variance is assignable to variance in the capacities measured by the Purdue assembly test. With the addition of the tweezer dexterity test, the amount of criterion variance accounted for becomes 51.8%, an increase of 11.4%. The further inclusion of the placing test raises this figure to 57.0%, an increase of 5.2%. Increments of 0.8% and 0.1%, respectively, would attend the addition to the battery of the finger dexterity and turning tests.

It appears therefore that, while the chance errors added to the battery by the successive tests have not completely outweighed the gains in forecasting efficiency, it would almost certainly be uneconomical of time and money to include the finger dexterity and turning tests in the final battery.

The three-test equation. Considering both forecasting efficiency and optimum utilization of testing time, a multiple regression equation, based upon the first three tests of Table 9, was developed. In *standard score form,* this equation reads

$$z_c = .374z_{pa} + .316z_{td} + .267z_p$$

where subscript, c, indicates the criterion and subscripts pa, td and p represent the Purdue assembly, tweezer dexterity and placing tests, respectively. Substituting the means and sigmas required to put this equation into the more convenient *raw score form,* yields the following:[9]

$$X_c = .414X_{pa} + .451X_{td} + .392X_p - 115.2$$

The standard error of estimate, in terms of the parameters of the assumed criterion scale, is 9.33.

The two-test equation. A second equation, based upon the Purdue assembly and tweezer dexterity tests, was developed for occasions when only limited time is available for testing applicants. Utilizing the same symbols and subscripts as above, this equation quantitatively relates the standard and raw scores of the respective measures as follows:

$$z_c = .469z_{pa} + .380z_{td}$$

and,

$$X_c = .545X_{pa} + .569_{td} - 76.2$$

The standard error of estimate of the equation in raw score form is 9.88. The

[9] The standard deviation of the assumed criterion scale (14.286) was divided by the multiple correlation coefficient (.757) before substitution, in order to correct the predictions of the raw score form of the equation for reduced dispersion (10). While this correction is not generally found in standard texts, its effect on the spread of the predicted criterion scores is so considerable as to make it imperative when the standard deviation of the predicted scores is used to evaluate the performance of applicants. The raw score form of the two-test equation is similarly corrected.

increase from 9.33 represents the loss in precision resulting from the elimination of the placing test. The loss is not substantial compared with the magnitude of the standard errors themselves. A similar conclusion was suggested above, when the addition of the placing test to the battery increased the amount of criterion variance accounted for, from 51.8% to 57.0%, an increment of 5.2%.

The standard error of estimate is a measure of the effectiveness of the battery when interest centers in predicting the performance of individual applicants. For example, if job performance were graded on the assumed criterion scale, the criterion score predicted by the two-test equation will be in error by 9.33 or less in roughly two-thirds of the cases. By the same token, a greater error will attend the predictions in one-third of the cases. When one considers the range of scores on the criterion scale, the errors for individual prediction are considerable. However, as Garrett (4) has stated,

> It may be argued . . . that in attempting to predict individual performance from test scores we are asking too much of our battery of tests—more than we have a right to expect. . . . From tables of life expectancy one can tell quite accurately how many men, now aged 30, will survive to age 50. But prediction of the life span of a given individual is a dubious undertaking.

Tiffin (18), working from the Taylor-Russell tables, comes to a similar conclusion. Both suggest the percent of proper placements as a more suitable measure of the effectiveness of tests in the employment situation. Accordingly, data from the present experiment will be analyzed from this point of view in a later section of this chapter.

The raw score forms of the two equations define the optimum use of the test scores in predicting the performance of applicants in the Training school. Any weighting of the tests, other than those indicated by the respective equations, would result in a loss in forecasting efficiency with respect to the criterion, for the experimental group. In order to reduce the computational labor involved in applying these equations to a mass testing situation, facilitating tables (10) were developed.

E. CONVERTING PREDICTIONS ON THE ASSUMED SCALE TO GRADES ON THE RATING SCALE

It was stated earlier that, instead of endeavoring to predict directly to the original platykurtic and negatively skewed rating distribution, an assumed criterion scale was constructed by selecting a mean of 50.000 and a standard deviation of 14.286 as parameters of the criterion distribution. It now becomes necessary to relate the predictions made by the raw score forms of the equations to the five grades of the original rating scale. This was accomplished by dividing the area under the normal curve defined by the assumed parameters into five parts, so that the percentage of the total area in each of the five segments corresponded successively to the percentages in each of the five rating categories. For example, 9.0% of the trainees were rated Poor. The point on the assumed scale, below which 9.0% of the people score is 1.34σ below the mean. Since the standard deviation was taken to be 14.286, this corresponds to a distance of 19.1 below the assumed mean of 50.0, or a score of 30.9. Thus, when test scores are substituted in the regression equation and a criterion score of 30.9 or less is obtained, this is equivalent to a prediction of Poor. A similar treatment applied to each step of the original scale, yielded four criterion scores which define the separation

between contiguous pairs of ratings.

Poor–Fair	30.9
Fair–Average	40.1
Average–Good	51.1
Good–Very Good	60.9

These results were used to locate the ratings at the top of the chart which will be described in the next section.

F. CHART FOR REPORTING TEST PERFORMANCE

In order to present test results to the employment office in a way which would depict the relative standing of the applicant in precise and readily comprehensible form, a chart similar to Figure 3, was devised. When an applicant is tested, raw scores are posted on the short horizontal lines at the right. The relative standing of the applicant on each test is then denoted by a red pencil mark at the appropriate point along the adjacent horizontal line in the body of the chart. Weighted scores from the facilitating tables are written to the right of the raw scores and added. Since the weighted scores in the tables were calculated in such a way as to absorb a large part of the constants (115.2 or 76.2), 20.0 is subtracted mentally from the sum and the result is posted on the short line to the right of the scale for the overall score. The corresponding point on the chart is checked and represents the statistically best prediction of the applicant's performance which can be made on the basis of the test results.

Thus the chart presents graphically the standing of an applicant on each of the tests as well as a prediction of her future job performance. Each of these items can be read on the scale of overall scores, which putatively represents units of equal ability; on the percentile scale, which shows the percentage of applicants scoring better than the one under

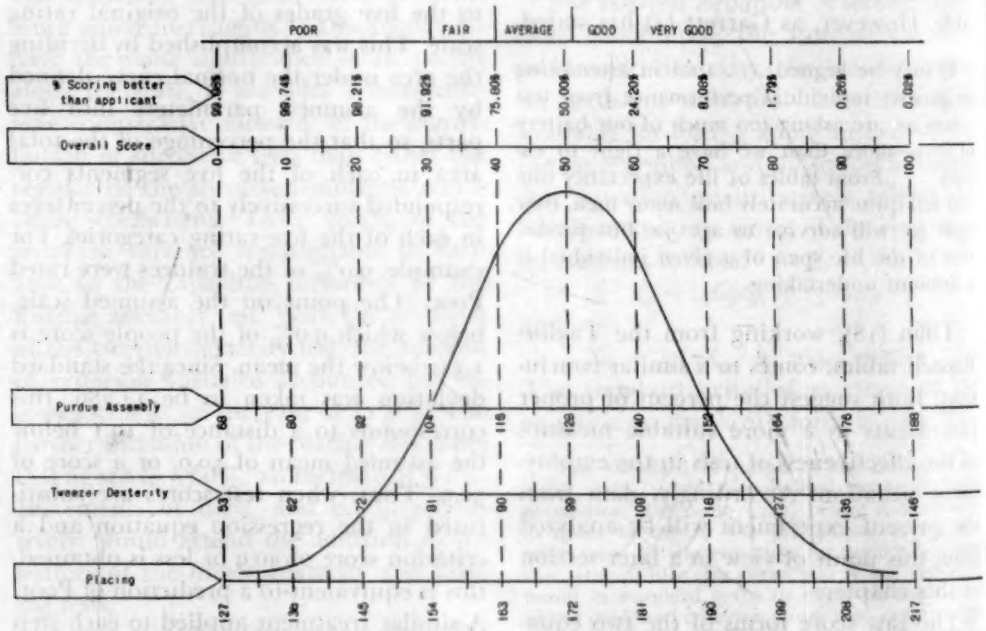


FIG. 3. Chart for Reporting Test Performance and Predicted Criterion Score to the Employment Office.

consideration; or, as a predicted job performance rating. The scale values of all measures were equated by locating each of the ten major divisions on the chart as a multiple of 0.7σ from the mean of each distribution. This fact makes it possible to regard all tests as if they were scored on the same numerical scale as the overall score. It would have been unnecessary, therefore, to have recorded the raw test scores on the chart, except that this facilitates marking the standing of an applicant on a given test. Needless to say, the normal curve, drawn in the body of the chart, applies to all scales.

G. PREDICTING FACTORY PERFORMANCE

It was implied above that the equation based upon the Purdue assembly, tweezer dexterity and placing tests is to be preferred, when time permits, to the one utilizing the scores on only the first two tests. This was the conclusion reached by the author upon completion of the experiment with performance in the Vestibule Training School as the criterion.

At a later time, however, when follow-up and termination ratings on subsequent performance in the factory were available, it became apparent that the two-test equation was superior in predicting performance with respect to these factory criteria. This discovery was of considerable practical importance since there is little point in adding a third test to the battery in order to account for an additional 5.2% of criterion variance in the training school if this procedure entails a loss in forecasting efficiency when the equation is applied to the factory. The analyses which were made to determine the reasons for, and magnitude of this discrepancy, involved the use of several criteria of factory performance as well as a considerable number of subjects who received the tests before and after the present study. However, in order to keep this report within manageable bounds and at the same time to separate the results which appear to have worthwhile, practical applications from those of lesser value, the following more limited evidence from the present experimental sample is offered. In this way conclusions may rest upon an empirical foundation without ambiguous references to unpublished studies.

To accomplish this, it is necessary to introduce a new criterion. At the time of the present experiment, it was the practice to send follow-up rating forms to the immediate supervisor of all applicants and trainees who were placed, after testing, on any one of the occupations then under investigation. The two most important items on the form were the information identifying the specific operation being performed by the individual and a summary rating on "Overall job performance, day in and day out." The scale consisted of the four grades, Poor, Fair, Good and Excellent, which have been in use for a number of years on the company's termination form. The form was sent to the factory when the operator had completed her tenth week on the job.

Of the 233 subjects in the experimental group, 122 (52.4%) were sent to the factory as mounters.[10] However, 63 of these left the company before or around the time when the follow-up ratings were sent out. By far the majority of these terminations occurred early in September and were effected by the individuals themselves for the purpose of resuming their education. Moreover, of the 59 forms which should have been returned, the net

[10] Some evidence for the effectiveness of placements from the Vestibule Training School is provided by the fact that no one in this group of 122 mounters was terminated as "Unsuited to the Work."

yield of this sample for the criterion under consideration was 35 ratings.

A breakdown of the 35 ratings showed that no one was rated Poor, in all probability reflecting in part the reluctance of supervisors to use this grade and in part the screening which occurred in the training school. Thirteen were rated Fair, eighteen, Good and four, Excellent. For the purpose of calculating biserial correlations, the 35 subjects were divided into two groups. The thirteen rated Fair were classified as "unsatisfactory" and the twenty-two rated Good and Excellent, as "satisfactory." Because of the sample size, the required means and standard deviations were calculated from the ungrouped data.

Two sets of predicted criterion scores were derived by substituting in each of the two regression equations, the raw scores obtained by the subjects on the prospective tests at the time of hiring. The resultant scores were regarded as predictions, on the assumed criterion scale, of the level of efficiency to be achieved by each subject as a mounter in the factory. The question now is, which equation has made the better predictions?

The second column of Table 10 provides the answer to this question as it relates to this particular group of mounters. Converting the correlations to coefficients of determination and expressing the result as a percent, the equation based on only two tests accounts for 35.4% of the criterion variance, whereas, the equation utilizing the scores on the three tests accounts for 18.6%.

While these statements are entirely correct when applied to the performance of the 35 mounters in question, they do not constitute a good basis for comparing the two equations. A glance at columns three and four of Table 10 will show that the factory group has a smaller spread on the criterion scale, with respect to the abilities measured by the tests and weighted by the respective equations than the original experimental group. This is to be expected in view of the screening which occurred in the Vestibule Training School. It is well known, however, that the spread of the scores has a substantial effect upon the size of a sample r. And, since the reduction in spread is different for the two equations, the two coefficients are differentially affected. The correlations must therefore be corrected for reduced dispersion before their relative worth can be determined. The corrected coefficients (14) are entered in column five of the table. Squaring the two corrected coefficients and multiplying by 100, the two equations may be compared more properly as follows: If the group of 35 mounters had been distributed over the criterion scale in the same way as the total applicant group, the two-test equation would have

TABLE 10

Comparison of the Two Regression Equations with Respect to Efficiency in Predicting the Performance of Mounters in the Factory*
(N = 35)

Equation	r_{bis}	$\sigma_{dist.}$ N = 35	$\sigma_{dist.}$ N = 233	r′	$t_{r'}$
2-test equation	.595	11.96	14.286	.678	3.16
3-test equation	.431	13.56	14.286	.477	2.22

* $\sigma_{dist.}$ for N = 35, equals the computed standard deviation of the predicted criterion scores of the 35 subjects. $\sigma_{dist.}$ for N = 233, is the standard deviation of the predicted criterion scores of the entire experimental group.

accounted for 46.0% of the criterion variance, whereas, the three-test equation would account for only 22.8%.

However, the interest of the reader who is concerned with the practical decision as to whether or not the data justifies the use of the tests in the employment office, goes beyond the correlations found with the particular subjects of the sample. It is well known that rather high correlations sometimes arise by chance, with small samples. Before much confidence can be placed in the results, therefore, it is necessary to determine the probability that the obtained correlations could have arisen by chance without the existence of a true relationship between the tests and performance. For this purpose, the standard error of a biserial r of zero was calculated for these data. It came out to be .2145. In deciding which set of correlations should be tested, the uncorrected or the corrected, it should be noted that the standard error of a biserial r does not take into account the reduced dispersion of the experimental group. If a correlation coefficient has been decreased by reduced dispersion, it may legitimately be corrected to a coefficient which describes the same amount of correlation for the population, in this case the applicant group. Therefore, the $t_{r_{bis}}$ entries in the last column of the table, provide the best estimate of the confidence which can be placed in each of these equations. For 33 degrees of freedom, t must equal 2.035 or 2.734 to be significant at the 5% and 1% levels, respectively. Tested in this way, the two-test equation is seen to be significant at far better than the 1% level, whereas, the three-test equation is significant somewhere between the 5% and 1% levels. Since practical interest centers in coefficients indicating a positive correlation, it may be said that there is less than one chance in 200 that the obtained correlation between the predictions of the two-test equation and job performance could have arisen by chance without the existence of a true relationship. There are less than 2.5 chances in 200 (i.e., 1 in 40) that a similar statement is true for the three-test equation. These statements are implicit in the definitions of the 1% and 5% levels, respectively.

It is worthy of mention, that the uncorrected r_{bis} for the two-test equation yields a t of 2.77 which, at 33 degrees of freedom, is also significant at the 1% level. The uncorrected coefficient for the three-test equation barely misses significance at the 5% level, with a t of 2.01. The fact that all four coefficients are probably attenuated by chance errors and the fact that neither the tests nor the performance ratings are perfectly reliable, lends further support to the testing of the corrected correlations. This is true in spite of the fact that a correlation corrected for the lack of perfect reliability of both measures would hardly be of practical value, unless one planned to lengthen the tests. Correlations corrected for chance errors in supervisory judgment would be of interest and, in all probability, substantially higher. There is little question, however, that confidence in the results is enhanced by the fact that both the corrected and the uncorrected coefficients for the two-test equation are significant at the 1% level.

Assuming that no other data were available and that the amount of confidence which can be placed in the predictions of factory performance should be a determining factor, there is little question that the two-test equation should be selected.[11] Therefore, only the results

[11] Since the b- and beta-coefficients in the equations turn upon relatively small differences in zero-order r's, the data are not adequate to determine whether or not predictions to the fac-

from the equation which weights the scores from the Purdue assembly and tweezer dexterity tests will be presented graphically in the next section.

H. GRAPHICAL PRESENTATION OF THE RESULTS

It has already been established that the equation utilizing the scores on the Purdue assembly and tweezer dexterity tests makes predictions which are significantly related to performance in both the training school and the factory. Although the graphical presentation of results will involve restating these relationships in less precise (though more palpable) form, several additional and important practical conclusions can be demonstrated more simply in this way.

Tiffin (18) uses the term, "selection ratio," to designate the percentage of the applicant group which must be placed on a particular job. If, from a group of one hundred applicants, one selects the forty who score highest on a test, he is operating with a selection ratio of 40%; whereas, if the highest seventy applicants are placed on the job, the selection ratio is 70%.

Figure 4 shows the effect of the selection ratio on the percent of operators who fall in the rating categories specified on the curves. The chart is based upon an array in which the predicted criterion scores for the 233 subjects were tabulated against the overall performance rating received in the school. Both the selection ratios and the percent satisfactory were calculated cumulatively from the high end of the test score distribution. The percentages designated as selection ratio have the same meaning as those at the top of the chart for reporting test performance (Figure 3). To facilitate comparisons between the two sketches, selection ratios increase from right to left in both cases. The alternate numeration on the abscissa represent the passing scores corresponding to the various selection ratios. These were derived from the assumed population parameters of the criterion distribution, and are directly comparable to the scores on the overall scale of Figure 3.

Curve (A) delineates the effect of variations in the selection ratio upon the percent of good and very good people in the group selected for the job. The fact that this curve intersects the vertical line corresponding to a selection ratio of 100% at an ordinate value of 46.8%, indicates that when the entire group of 233 subjects was hired, 109 (46.9%) were rated Good and Very Good in the school. However, a substantially greater percentage of good and very good trainees could have been sent to the school if the tests had been used in selection, as demonstrated by the rapid rise in the curve as the selection ratio becomes smaller. It was noted earlier, for example, that 52.4% of the trainees were assigned to the factory as mounters. If this group had been selected for the school by tests, 62.9% would have been rated Good and Very Good, an increase of 16.1%. If only 40% of the total group had been selected on the basis of the tests, the percent of good and very good trainees would have risen to 70.5%, an increment of 23.7%.

The attainable improvements in the percent of very good people are even more substantial. Thus 22.3% of the entire group were rated Very Good. From line (B) it is apparent that this figure could have been increased to 34.7% if the 122 (52.4%) trainees who were finally assigned to the factory as mounters had been selected by the tests. If only

tory criterion could be improved still further by revising the weights assigned by the school criterion.

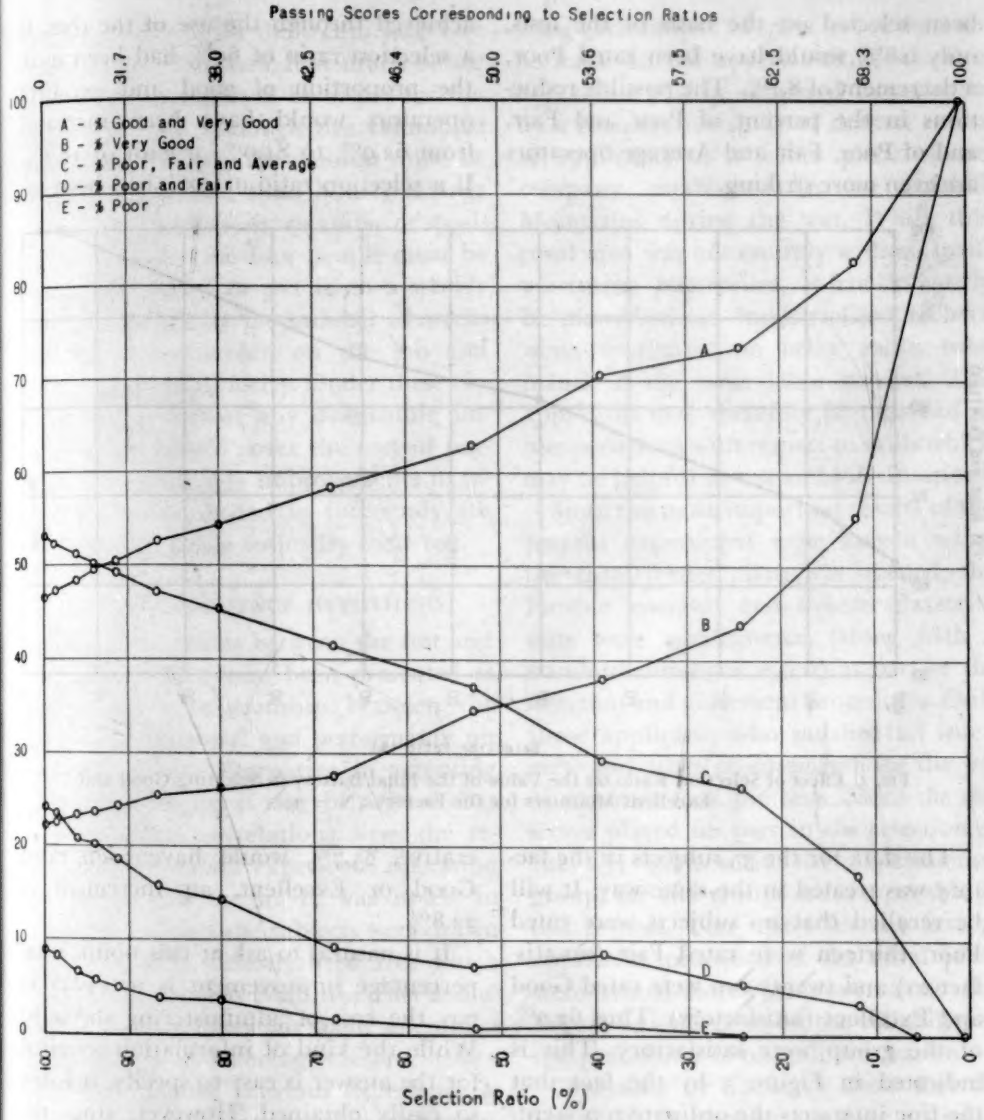


FIG. 4. Effect of Selection Ratio on the Percent of Trainees in Indicated Rating Groups, When Selected by the Final Battery. (N = 233)

30% had been selected, 42% would have been rated Very Good, an increase of approximately 20%. The effect of smaller selection ratios may be similarly read from the graph.

Obversely, the lines representing the three lowest ratings show the extent to which the percent of poorer trainees in the selected group can be decreased by reductions in the selection ratio. Thus, while 9.0% of the total group were rated Poor, line (E) demonstrates that this could have been reduced to 2.0% by operating with a selection ratio of 70%. Furthermore, if the 52.4% who were assigned from the school as mounters had

been selected on the basis of the tests, only 0.8% would have been rated Poor, a decrement of 8.1%. The possible reductions in the percent of Poor and Fair, and of Poor, Fair and Average operators are even more striking.

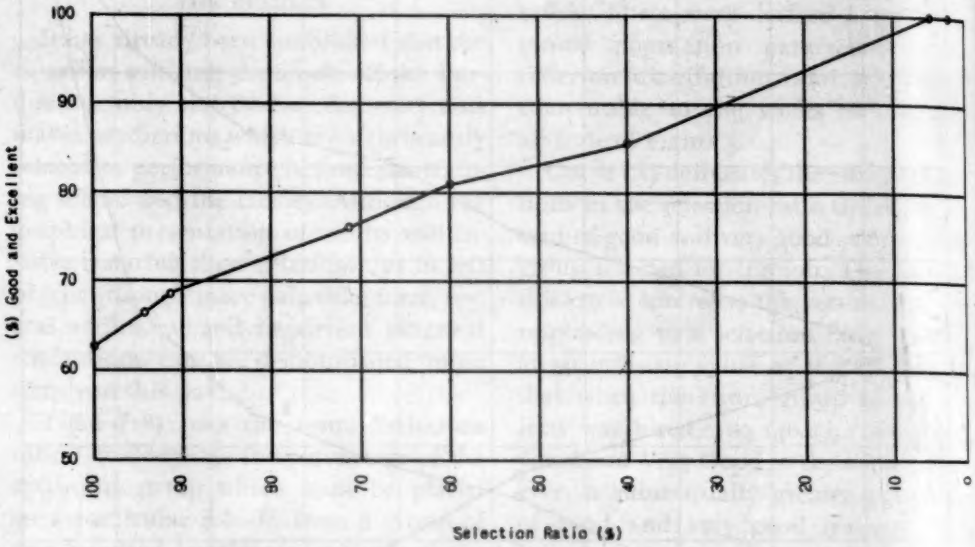


FIG. 5. Effect of Selection Ratio on the Value of the Final Battery in Selecting Good and Excellent Mounters for the Factory. (N = 35)

The data for the 35 subjects in the factory was treated in the same way. It will be recalled that no subjects were rated Poor, thirteen were rated Fair (unsatisfactory) and twenty-two were rated Good and Excellent (satisfactory). Thus 62.9% of the group were satisfactory. This is indicated in Figure 5 by the fact that the line intersects the ordinate representing a selection ratio of 100%, at 62.9%. This figure is undoubtedly higher than it would have been if the subjects had not been screened through the Vestibule Training School. We have seen the effect of reduced dispersion on the correlation coefficient. The relationship presented in the graph is similarly affected. Despite this limitation, the line shows that substantial improvements could have been achieved through the use of the tests. If a selection ratio of 60% had been used, the proportion of good and excellent operators would have been increased from 62.9% to 81.0%, a gain of 18.1%. If a selection ratio of 40% had been operative, 85.7% would have been rated Good or Excellent, an increment of 22.8%.

It is natural to ask at this point, what percentage improvement is necessary to pay the cost of administering the tests? While the kind of information required for the answer is easy to specify, it is not so easily obtained. However, since the tests can be administered on a group basis, up to ten or twelve applicants can be tested conveniently at one time. If one is operating with a selection ratio of 40%, ten applicants must be tested for every four placed on mounting jobs. Since the tests can be administered in less than twenty minutes, it is doubtful, under any circumstances, whether the total cost of testing this group would be greater than

one hour's salary for the technician. The question now becomes, how much better must these four people be to pay for the testing of ten? Although this cannot be answered precisely, the amount must be small indeed. The total hourly savings through superiority in quantity or quality of work for the four people must be multiplied by 40 to put it on a weekly basis, and then by the number of weeks the operators remain on the job and demonstrate superiority. Under these circumstances, almost any discernible improvement should cover the cost of testing. The attainable improvements in selection, demonstrated in this study are certainly adequate to justify their use.

I. THE EXPERIENCE HYPOTHESIS

The correlations between the test and the two criteria have been presented as indicating a relationship between the aptitudes measured and performance on the job. As an alternative to accepting this major thesis, it may be argued that the empirical correlations were the result of differential experience operating as a spurious factor. It was noted in Chapter IV, that the subjects were drawn from a highly industrialized area which includes numerous establishments manufacturing miscellaneous small assemblies and several producing vacuum tubes. Trainees having previous experience as mounters were eliminated with reasonable certainty through the experimental procedures. Rehires to the company were also excluded regardless of position held. However, these precautions do not necessarily provide adequate assurance that *all* experience which might spuriously inflate empirical correlations has been controlled.

In order to avoid the difficulties and uncertainties which would attend evaluating each subject's report of her own experience on the application blank, this hypothesis will be rendered improbable by a comparative study of the test scores of applicants to a feeder plant of the company, established in the Catskill Mountains during the war. While this rural area was not entirely without manufacturing enterprises, it could hardly be described as industrialized. There were, moreover, no other radio tube plants in the same labor market. The applicants may therefore be regarded as inexperienced with respect to skills which may be helpful in test and job situations.

Since the more important results of the present experiment were known when the rural feeder plant was opened, the Purdue assembly and tweezer dexterity tests were administered (along with a standard intelligence test) as part of the selection and placement procedures. Only those applicants who satisfied all interview and other requirements for the job of mounter took the tests. Since the test scores played no part in the selection of the 233 subjects in the experimental group, the two groups are comparable in this respect. Test scores for the first 78 applicants processed at the rural plant were treated shortly after operations were begun, as a check upon the reliability of the procedures of test administration.

The data in Table 11 show that the scores of the rural group have a slightly higher mean and a slightly greater dispersion on the Purdue assembly and tweezer dexterity tests than those of the experimental group. It will be noted that all differences, though small, are in a direction contrary to that which would be expected on the basis of the experience hypothesis. However, since the critical ratios range from 0.16 to 0.84, the differences between the two groups could easily

TABLE 11

Comparative Test Performance of the Experimental Group (N = 233) and a Group of Applicants to a Rural Plant (N = 78)

Tests	Statistics	Rural	Experimental	Diff	σ_D	Critical Ratio
PA	M	129.07	127.02	1.15	2.27	0.51
	σ	17.48	17.05	0.43	1.16	0.27
TD	M	100.85	99.38	1.45	1.57	0.84
	σ	13.41	13.21	0.20	0.86	0.16

have arisen by chance. None even approach statistical significance.[12]

These comparisons by themselves are not sufficiently complete to be conclusive. It is possible for two distributions of vastly different shape to yield comparable means and standard deviations. The chi-square test was therefore applied to the four distributions to detect any significant deviations from normality. Theoretical frequencies, calculated for each cell from the respective means and standard deviations, were added successively from either end of the distributions. Each time the total just reached or exceeded five, a single cell was constituted. Because of the smaller number of subjects in the rural group, a greater number of step intervals in the tails were combined. The calculated values of chi-square and the corresponding degrees of freedom are entered in Table 12. Significance levels were derived by interpolation between published values (5) and show the probability that chi-squares as great or greater than those obtained could have arisen by chance. For example, deviations from normality as great as, or greater than, those of the experimental group on the Purdue assembly test would arise by chance 41% of the time when the true shape of the distribution is normal. Since none of the figures even approaches the 5% level, we may conclude that the distributions are not significantly different from normal.

If differential experience had been pres-

[12] Large sample statistics were used in testing the reliability of the differences between the two groups. Since none of the differences approach significance, the more refined tests were considered unnecessary. Prior to publication, however, these assumptions were checked using population variances and taking into account the unequal sizes of the samples (1). Application of the t-test to the differences between the two means yielded values of .51 and .83 for the Purdue assembly and tweezer dexterity tests respectively, as compared with .51 and .84 for the critical ratios. Moreover, the ratio of the larger to the smaller variances (F) became 1.06 for the Purdue assembly and 1.04 for the tweezer dexterity test. None of these values is significant at the 5% level for the degrees of freedom involved.

TABLE 12

The Chi-Square Test of Normality Applied to Score Distributions of Rural and Experimental Subjects

Tests	Group	χ^2	Degrees Freedom	Level of Significance
Purdue Assembly	Experimental	16.74	16	41%
	Rural	9.20	11	60%
Tweezer Dexterity	Experimental	11.55	17	82%
	Rural	13.89	11	24%

ent to the extent that it materially affected test performance, both the means and standard deviations of the experimental group would have been larger than those of the rural group. The obtained differences, however, while in the opposite direction, were so slight as to be insignificant. The fact that the chi-square test showed no significant deviations from normality provides additional evidence against the experience hypothesis, since if the aptitudes measured are normally distributed, any factor differentially affecting test performance would lead to deviations from normality.

Thus the hypothesis that the empirical correlations arose spuriously from the effect of differential experience on both test and job performance appears highly improbable.

The work experience of the 35 subjects included in the factory follow-up, on the other hand, was checked on an individual basis. Of this group, 22 had no previous work experience. Nine of the remaining 13 had no industrial experience. This group included five salesgirls, two domestic workers, one waitress and one assistant to a beautician. Only four had had any industrial experience at the time of hiring. Two of these had performed an unspecified sewing operation on handbags for a number of years. However, one, with a predicted criterion score of 55.6 (on scale for overall scores, Figure 3), was rated Good in the factory, and the other, whose score was 38.1, was rated Fair. The third described her work as "sewing machine operator" without specifying the product. She received a predicted score of 56.2 and was rated as Excellent. However, she had left the sewing job after trying it for only two weeks. The fourth had worked eight months packing lipsticks. She received a rating of Good in the factory and a predicted criterion score of 69.7. The fact that one applicant with more than five years' experience sewing handbags scored below average on the tests and received a fair rating in the factory, suggests that success in a sewing operation is not necessarily helpful to mounters in either the test or the job situation. Packing lipsticks, moreover, is hardly comparable to mounting. Thus none of the 35 operators had previous industrial experience with tweezers or in the assembly of small and delicate parts.

The fact that no direct experience in mounting was found in this sample serves as a check on the reliability of the experimental procedures for removing the spurious effect of this factor. The weight of this evidence is enhanced when one considers that the 35 subjects rated constitute a sample from the 59 cases (p. 29f) where experience, if any, was most likely to be found. It will be recalled that 122 of the original 233 trainees were assigned from the school as mounters. Since skill in this operation was at a premium during this period, one may safely assume that the remaining 111 subjects were without helpful experience. Of the 122 assigned to mounting, 63 left the company before the ratings were made. The bulk of these subjects were vacation workers from the schools and colleges. Thus, any subject who had returned to mounting, after previous experience with it, would more likely be found among the remaining 59 operators than in any of the other groups. It is probable, therefore, that a detailed check of the application blanks of the 233 subjects would have revealed no experienced mounters and that the experimental controls on this factor were adequate.

VI. Summary and Conclusions

The assembly of small radio tube parts by the process of resistance welding was seen to require better than average ability in the manipulation of small and delicate parts with fingers and tweezers. A detailed description of the operation served to indicate the high degree of control which must be exercised over the movements of the hands and fingers in positioning the parts between the two parallel electrode contact surfaces. Although other aptitudes and worker characteristics appeared to be involved, the present study was confined to the evaluation of five standard manipulation tests, namely, the Minnesota Rate of Manipulation Tests, Turning and Placing, the O'Connor Finger Dexterity and Tweezer Dexterity tests, and the assembly subtest of the Purdue Pegboard. These tests were administered on a time limit basis to 233 prospective mounters at the time of hiring. The scores achieved were later correlated with a criterion of performance in the Vestibule Training School, consisting of the pooled judgments of the school supervisor and the staff of instructors.

Several types of data were analyzed in order to evaluate the criterion of performance in the school. The intercorrelations among three sets of ranks from a previous experiment, showed that the raters were capable of making reliable judgments of trainee performance on the basis of recorded information such as production test scores, number of defective units produced and other notations made in the course of training. Overall rankings, moreover, were found to correlate highly with separate judgments of quantity and quality. The correlations between criterion ratings and production test scores were taken to indicate both the reliability and validity of the criterion. The fact that production test scores were reasonably reliable measures of the aptitudes involved in the performance of each operation was interpreted as further evidence of the validity of the overall ratings based upon them.

Correlations between the criterion and the respective tests ranged from .482 to .636, far exceeding the requirements for significance at the 1% level for 231 degrees of freedom. Application of the Wherry-Doolittle Test Selection Method showed that, while the shrunken multiple correlation coefficient was a maximum when the battery included all five tests, the turning and finger dexterity tests contributed little to forecasting efficiency. Two multiple regression equations were then derived. The three-test equation, utilizing scores from the Purdue assembly, tweezer dexterity and placing tests, was found to account for 57.0% of the criterion variance. The two-test equation, which weighted scores on only the first two of these tests, accounted for 51.8% of the criterion variance, and was recommended for occasions when only limited time was available for testing.

In order to separate the results which appear to have worthwhile practical applications from those of limited value, the predictions made by the two equations were further validated against a criterion of factory performance. Even though criterion data were available for only 35 members of the original group of 233 trainees, the data were adequate to demonstrate that the predictions of the two-test equation are significantly related to factory performance at better than the 1% level. The forecasts of the three-test equation, on the other hand, barely missed significance at the 5% level prior

to correction for reduced dispersion and somewhat exceeded this minimum requirement when corrected. On the basis of these results, the two-test equation was recommended for application in the employment office. There appeared to be little point in including the placing test in the battery to account for an additional 5.2% of criterion variance in the training school, if the procedure entails a loss in forecasting efficiency with respect to factory performance.

The predictions of the two-test equation were then analyzed graphically to show the improvements in the percent of superior employees who would have been hired, operating with various selection ratios. It was demonstrated that, with selection ratios of moderate size, substantial increases were attainable in the percent of superior employees placed in both the training school and the factory.

The comparisons of the scores of the experimental subjects with a group of applicants to a rural feeder plant served as a check on both the procedures for eliminating experienced mounters and the possible spurious effect of other industrial training. The absence of any significant difference between the two groups and the fact that none of the score distributions deviated significantly from normal, were interpreted as rendering the experience hypothesis highly improbable. Some additional confidence in these conclusions may be derived from the fact that the empirical differences between the means and standard deviations, though slight, were in a direction opposite to that which would be expected on the basis of the experience hypothesis. An individual check on the work experience of the 35 mounters who were rated in the factory revealed that none had had previous industrial experience with tweezers or in the assembly of small and delicate parts.

A chart was developed for presenting test results to the employment office in a precise and readily comprehensible form. In addition, it provides the interviewers with the selection ratio implied by the acceptance of any individual applicant. During the period immediately following the installation of tests, the chart served as a training device, showing not only the magnitude of individual differences, but the nature of their distribution.

Viewing the results in perspective, it would be naive to pretend that two dexterity tests will solve all selection problems for the occupation of radio tube mounter. Valuable, as they have been shown to be, there remains a sizeable amount of residual variance to be explained. Available measures of intelligence and temperament, such as those employed in the study by Forlano and Kirkpatrick (2), should be validated with larger samples. The investigation of individual differences in perceptual and visual aptitudes may also prove fruitful. The possibility of devising new and perhaps better manipulation tests must not be overlooked. Finally, the development of techniques to increase the percent of mounters who make full use of their aptitudes in the daily performance of their jobs and who derive a reasonable amount of satisfaction therefrom is an even more challenging field for future research.